Psychological Bulletin

VOLUME 58, 1961

ARTHUR C. HOFFMAN, Production Manager
HELEN ORR, Promotion Manager

PUBLISHED BIMONTHLY BY
THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
1333 SIXTEENTH STREET N.W.
WASHINGTON 6, D. C.

ACKNOWLEDGEMENT

The following individuals have served as consulting editors in addition to the regular panel of consultants during the past year:

F. A. Beach
W. J. Brogden
D. E. Byrne
E. J. Capaldi
C. N. Cofer
K. M. Dallenbach
James Deese
Benjamin Fruchter
C. F. Hereford
Ira Iscoe
L. A. Jeffress
J. J. Jenkins
R. K. Lindsay
Quinn McNemar
R. E. Morin
R. W. Moulton
Gardner Murphy
M. E. Noble
D. A. Riley
T. R. Sarbin
Martha A. Wilson
Philip Worchel
R. K. Young

To them the Editor expresses his thanks for helping in the task of determining suitability of material submitted for publication in this journal.

Harry Helson, Editor

CONTENTS OF VOLUME 58

America's Psychologists

A Survey of a Growing Profession

By Kenneth E. Clark

University of Minnesota

A report of a study of American psychologists, sponsored by the APA Policy and Planning Board and supported by the National Science Foundation, which provides a clearer view of psychology in the mid-twentieth century by describing the people who are active in the field, and the nature of their activities. Some have been outstanding in research productivity. What are they like? How do they differ from their less productive colleagues? Are there major differences among psychologists in, say, experimental psychology and those in, say, industrial psychology? To answer such questions, Dr. Clark and his collaborators have studied the undergraduate education, family backgrounds, types of jobs held, and attitudes and values of different groups of psychologists. How many persons in the United States are engaged in predominantly psychological work? Are recent recipients of the Ph.D. similar to or different from those who received the degree 10 or 20 years ago? Where are psychologists employed? What do they read? These are samples of the questions that are discussed and on which substantial amounts of factual data are given in the pages of this report.

Price, $1.00

★

American Psychological Association
1333 Sixteenth St., N.W.
Washington 6, D.C.

APA Members and Journal Subscribers—Are you going to move?

If you move—

your journals will not follow you from your old address to your new one

When you move—

notify the APA Subscription Office

Formerly, journals that could not be delivered because subscribers had not notified the APA of a new address were reclaimed by the APA, and the journal was remailed to the subscriber at his new address. This was always expensive. Recent changes in the postal laws and regulations have made the expense prohibitive. Undeliverable copies are now destroyed by the Post Office. Subscribers who do not receive a journal because of an address change are charged the regular single issue price for a replacement copy.

So—when you move—

Notify the postmaster at your old address and guarantee that you will pay the forwarding postage.

Notify the APA Subscription Office as early as possible—by at least the tenth of the month preceding the month when the change should take effect.

AMERICAN PSYCHOLOGICAL ASSOCIATION
1333 Sixteenth Street N.W.
Washington 6, D.C.

VOL. 58, NO. 1 JANUARY 1961

Psychological Bulletin

TECHNIQUES FOR THE STUDY OF GROUP STRUCTURE AND BEHAVIOR:

II. EMPIRICAL STUDIES OF THE EFFECTS OF STRUCTURE IN SMALL GROUPS[1]

MURRAY GLANZER[2] AND ROBERT GLASER

American Institute for Research and University of Pittsburgh

An earlier paper (Glanzer & Glaser, 1959) reviewed techniques for analyzing the structure of groups that had been permitted to form their own pattern of interaction. This paper reviews laboratory studies in which experimenters imposed different structures on groups and measured the effect of the structures on performance.

The laboratory studies focus on communication structure. A communication structure is a set of positions with specified communication channels. Between any two positions, there may be a two-way channel, a one-way channel, or none at all. A channel is essentially the probability that a message can pass in a given direction between two positions. It may be defined more generally as the probability, p_{ab}, that a message can get from Position *a* to Position *b*. This is *not* the probability that *a* will try to send to *b*. It is, rather, the probability of his getting a message through if he tries to send one. In most of the structures studied, the channels are symmetric i.e., $p_{ab}=p_{ba}$ and the channels are either available or not, i.e., $p_{ab}=0$ or 1.

The studies are grouped in the following sections: The Initial Work, Variations and Further Analysis of the Basic Design, Testing the Limits of the Basic Design, Mathematical Analysis, Emphasis on Distribution of Functions in the Simulated Team, and Emphasis on Feedback and Learning. Tables summarize the main findings for the studies reviewed, often including more details than those covered in the text. The tables introduce a number of necessary simplifications. When an investigator employed several closely related measures, e.g., group morale, job satisfaction, status evaluation, findings on only one are included. Findings not presented in a form that permits evaluation are omitted. Findings on the effect of trials or learning, a statistically significant variable in almost all types of groups, are omitted unless especially rele-

[1] Prepared under Contract Nonr 2551(00), between the Office of Naval Research, Psychological Sciences Division, Personnel and Training Branch and the American Institute for Research, as part of a research project on team training and performance. The authors wish to thank Alex Bavelas, Harold Guetzkow, Robert L. Hall, John T. Lanzetta, Seymour Rosenberg, Marvin E. Shaw, and Gerald Shure, who read, and commented on, earlier drafts of the paper.

[2] Now at Walter Reed Army Institute of Research.

vant. In order to permit comparison between studies in both the tables and the text, the same terms will be used throughout for a set of related measures, even when this departs from the investigator's usage. For example, morale will refer to a variety of measures concerned with satisfaction in the experimental situation. For the same reason, although different names are used in various studies for the same network, one name will be used throughout this review.

The Initial Work

The area of communication structure was opened up 12 years ago by Bavelas (1948) with a discussion of mathematical aspects of group structure. The paper is Lewinian in tone, using the terminology of boundary, region, etc. The Lewinian boundary, however, is translated into the link or channel. This translation is of major importance for all the work that follows. Bavelas then builds up a set of assumptions and definitions concerning a collection of cells. He defines cell boundary, region, open cell, closed cell, region boundary, chain, chain length, structure, cell distance, cell-region distance, etc., and considers the factors that cause each measure to vary, deriving theorems concerning the following: the limits of the values for the various distances and other measures, the relation between the distance measures and the spread of a change of state in the structure, and, characteristics of pathways within the structure. Bavelas then shows how the various distances change as a function of structure types (e.g., organizations with varying degrees of horizontal coordination) and as a function of an increase in the number of levels in the organization. He also discusses the role of special positions such as liaison positions and possible applications of his approach.

The many provocative points raised in the paper were not directly followed by experimental work. Experimental work was set off by a second, much simpler paper (Bavelas, 1950), which differs markedly from the first. The Lewinian tone has disappeared. For example, regions within structures are not mentioned. Bavelas now discusses a few simpler concepts which readily generate experimental situations. Complex concepts such as inner and outer regions and chains of connecting cells do not appear again in the work in this area. The only concepts that survive from the first paper are those of links and distances. The focus of the discussion changes, moreover, from the larger *in situ* group, e.g., an industrial organization, to the small laboratory group.

In the second paper, Bavelas introduces the communication networks which were to become standard experimental arrangements. The channels of these networks are all two-way channels: a channel from *a* to *b* is also a channel from *b* to *a*. He also introduces the index of relative centrality to describe the structures. The index of the relative centrality (of Position *x*) is the ratio of the sum of the minimal distances of all positions to all others over the sum of the minimal distances of Position *x* to all others, or

$$C(x) = \frac{\sum_{x} \sum_{y} \mathrm{d}_{xy}}{\sum_{y} \mathrm{d}_{xy}} \qquad [1]$$

where d_{xy} is the minimal distance between x and y. Many of the investigations of group structure published after this paper focus on this measure. In subsequent studies $C(x)$ is also summed over all positions, x, in the network to give net centrality.

The main question Bavelas now asks is the following: Is it possible that

> among several communication patterns, all logically adequate for the successful completion of a specified task, one gives significantly better performance than another (p. 726)?

In answer to this question, he describes experimental results obtained by S. L. Smith (unpublished report) and Leavitt (1951) who use an experimental arrangement that is the model for the majority of the subsequent studies.

Five subjects were each given a list of symbols. Their task was to discover which symbol they all had in common. The physical setting was arranged so that some group members could send messages to each other, other group members could not. Smith's and Leavitt's subjects sat around a table partitioned into five sections with slots, some of which were open to allow notes to pass between sections. The pattern of open slots determined the communication pattern or structure. The subjects were free to use the open communication channels in any way they wished. They were not told the structure of the network.

The group's task required two main steps: distribution of individual information so that some or all members had all the necessary information, and determination of the common symbol. The task was completed when all subjects gave the answer. Smith imposed two communication structures: Circle and Chain (see Figure 1) and finds that structure affects group performance. The Chain is more efficient than the Circle. The performance ascribed to individuals is related to their positions. The central positions are most frequently seen as leaders. Table 1 indicates the type of data analysis carried out in the network studies. The two main independent variables are the patterns as units (Circle versus Chain) and the positions within a given pattern (*a*, *b*, *c*, *d*, and *e*.)

Leavitt (1951) used the same physical arrangement and problem as Smith did, but with four structures: Circle, Chain, Wheel, and Y (see Figure 1). His main positive findings are that the Wheel, Y, Chain, and Circle (most centralized to least centralized) rank in descending order (best to least) with respect to the following: (*a*) speed of development of organization for problem handling (the Wheel, Y, and Chain were, moreover, stable once they developed their organization. The Circle was inconsistent, i.e., problem solving procedure never became fixed); (*b*) agreement on who the group leaders were; and (*c*) satisfaction with the group. The ordering on these characteristics correlates perfectly with the ordering of the values of the net centrality index, $\Sigma_x C(x)$.

During the course of 15 trials, all the structures showed learning, reducing the time to complete trials. The networks did not, however, differ clearly from each other in speed or in learning rate. Leavitt asserts that the Circle used more messages and made more errors than the other networks. The interpretation of the data, however, is unclear since the analyses are based on a selection of the data, e.g., number of messages on successful trials.

Leavitt, in analyzing the effects of position within a network (see Table 2) finds that the most central position sends the most messages and the least central, the fewest. Subjects at

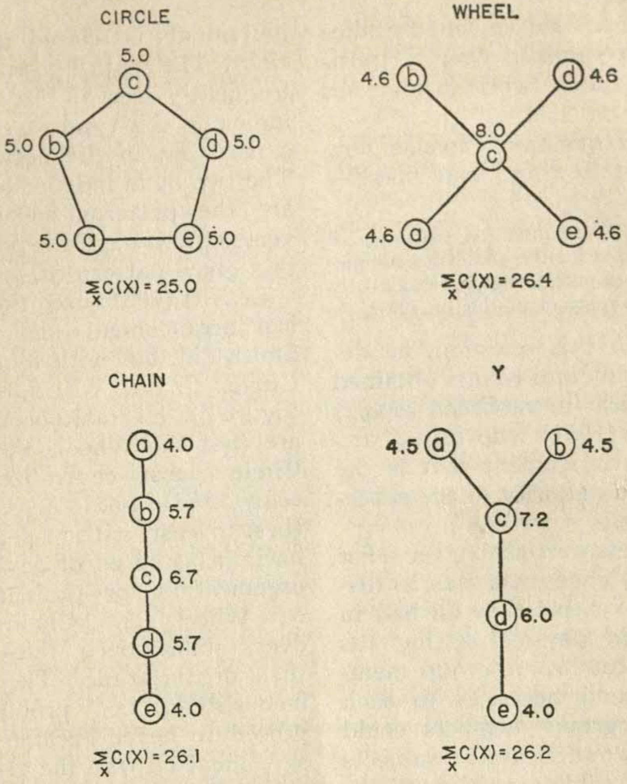


FIG. 1. Five-man networks used by Smith and Leavitt with the relative centrality index of each position and net centrality, $\sum_x C(X)$.

the central position, moreover, enjoy their jobs more than those at peripheral positions. The relation of centrality to number of messages is to be expected, since the central positions had to serve as relays for messages from the peripheral members. Concerning the relation between position and morale, Leavitt (1951) offers the following explanation:

> In our culture, in which needs for autonomy, recognition, and achievement are strong, it is to be expected that positions which limit independence of action (peripheral positions) would be unsatisfying (p. 48).

The dependent variables of the

TABLE 1

SUMMARY OF SMITH'S DATA FROM BAVELAS (1950)

Patterns	Average total errors	Average incorrect completions	Frequency of occurrence of recognized leader at position				
			a	*b*	*c*	*d*	*e*
Circle	14.0	5.0	1	2	3	2	3
Chain	7.0	1.5	0	1	14	3	0

Smith and the Leavitt studies are the major concern of the subsequent network studies. The variables fall into four main classes: (*a*) efficiency—number of errors, correct completions, speed of solution, number of messages; (*b*) leadership—positions named as leader, agreement about leader; (*c*) morale—rating of group, rating of self; and (*d*) organization—consistency, type. These dependent variables and the two main independent variables, group structures and individual position within group structures, form the basic framework of the network studies.

Variations and Further Analysis of the Basic Design

The work of Bavelas, Smith, and Leavitt proliferated into an abundance of network studies. The first of these was a study by Heise and Miller (1951), introducing the following variations in the original procedures: (*a*) Communication took place over an intercom system. The subjects could, therefore, listen or speak simultaneously to as many of the other subjects as the network permitted. (*b*) Communication content was highly restricted. The subjects could only relay the words on a given list. (*c*) The communication network included one-way as well as two-way channels. The five three-man structures used are presented in Figure 2. (*d*) Intensity of noise was varied over the networks.

TABLE 2

Summary of Leavitt's (1951) Findings for Structurally Distinct Positions

Pattern	Position	Mean number of messages sent	Mean job enjoyment
Circle	*a, b, c, d,* & *e*	83.8	65.6
Chain	*a* & *e*	26.2	34.5
	b & *d*	71.3	76.2
	c	82.4	78.0
Wheel	*a, b, d,* & *e*	28.1	31.2
	c	102.8	97.0
Y	*a* & *b*	25.9	47.5
	c	79.8	95.0
	d	63.8	71.0
	e	25.6	31.0

Using a task in which the subjects had to reconstruct a master list of words on the basis of incomplete lists, Heise and Miller find that: (*a*) As the signal-to-noise ratio in the intercom channels was lowered, the number of words spoken, errors, and the time required to complete the task increased for all networks. (*b*) With increased noise, the differences between networks became more pronounced. Generally, inefficiency of performance, measured by either the number of words spoken or the time required to finish the task, increased from Pattern 1 to 5 (in Figure 2). A second task in which the subjects had to reconstruct a 25-word sentence based on parts given to each of them,

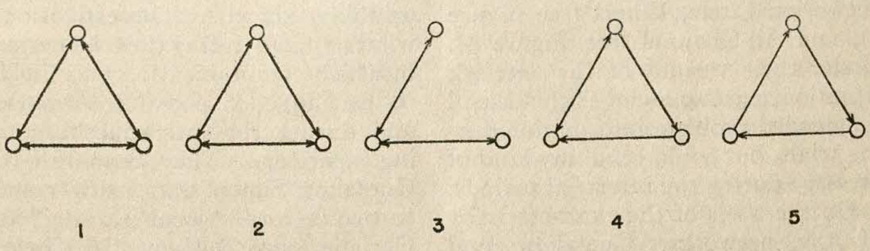


Fig. 2. Three-man networks used by Heise and Miller.

gave similar results, except that Networks 2 and 3 were somewhat more efficient than Network 1 at the high noise levels. When, however, the subjects were given anagram problems in which communication between the subjects was not necessary, the results were as follows: intense noise decreased the number of words spoken; there was no systematic difference between the efficiency of the various nets.

Aside from its introduction of a greater variety of channel arrangements, the main contribution of the study is probably that it demonstrates that no network is best in all situations. The efficiency of a structure depends on the characteristics of the task. Thus, in one of the first network studies, the complex interactions that will mar the apparent simplicity of the early findings appear.

Guetzkow and Simon (1955) introduced the distinction between two classes of behavior in the network: direct problem solving behavior, such as relaying information and asking questions; and organizational behavior, such as assigning of roles and functions to team members. They hypothesize that communication restrictions affect only the ability of the group to organize; once the group is organized, however, the different structures are equally efficient in solving the problems. To test their hypothesis, they used three five-man networks: Circle, Wheel (see Figure 1), and All-Channel (see Figure 4). Under their variant of the network situation, a group member could send only coded problem information during trials, but could send any kind of message during the intertrial periods.

On the basis of the characteristics of the networks, Guetzkow and Simon predict that the Wheel should be highest in efficiency, the All-Channel intermediate, and the Circle lowest.

> The Wheel groups would have the least difficulty, for they have no channels to eliminate, no relays to establish, and already have one person occupying a dominant position in the net. The All-Channel groups would have the next grade of difficulty, since the elimination of excess channels and the evolution of one person as solution-former are both required, yet relays need not be established. The Circle groups should have the most difficulty, for they need both to establish relays and to evolve an asymmetrical arrangement among the positions. They also must do some eliminating of unneeded channels, although this last requirement is minimal (p. 240).

Their findings on speed of problem solution (which also agree with Leavitt's contention concerning the Wheel and the Circle) bear out this prediction.

They cite the following as evidence that the structures affect organizational efficiency: The interaction patterns were most stable (same channels consistently used) in the Wheel and least stable in the All-Channel; the greatest degree of differentiation of function is found in the Wheel, the least in the Circle. They show, furthermore, that if only the stable groups of each network are compared, then there are no longer differences in the speed in problem solution. They cite that finding as evidence that the communication restriction does not affect the problem solving directly.

Guetzkow and Dill (1957) follow up this study with an investigation of what happens during the trial periods, in which communication was limited to exchange of coded information, and during the intertrial "organizing" periods. They reanalyze the Guetzkow-Simon data with respect to two factors—"local learning" (see Christie, Luce, & Macy, 1952, below) which occurs during the trials, and "planning mechanism" which functions during the intertrial period—and conclude that All-Channel shows the most planning activity while Wheel shows the least, presumably because its organization is dictated by the communication net. They furthermore note that the Circle network is handicapped in organizing itself during the intertrial period by the network restrictions, whereas the All-Channel structure does not seem to have this difficulty.

In order to explore this point, Guetzkow and Dill obtained new data by running groups of subjects under an alternating structure condition. During the task trials, the groups were run as Circles. During the intertrial periods, all communication restrictions were removed by opening the barred channels, giving an All-Channel net. These new experimental groups are called Circle–All-Channel. Guetzkow and Dill (1957) hypothesize that

> task performance in a restricted net will be equal to that in an unrestricted net, if the restrictions are removed during the intertrial period so that a relay system may be organized (p. 191).

An analysis of task trial times failed to support the hypothesis. Circle–All-Channel groups do not differ in performance time from the Circle groups in the earlier experiment. All-Channel groups were, moreover, significantly faster than Circle–All-Channel groups. The main contribution of the two studies above is their suggestion concerning the ways in which communication structure impedes the group's attempts to organize itself for its work.

Goldberg (1955) brings to the network study a new task, the unstructured group decision task, and a new dependent variable, influence (or, more precisely, "influenceability"). He hypothesizes that in group decisions, central positions in a network would be influenced less than peripheral positions. He placed subjects in the five-man Wheel, Y, and Chain and showed them a card bearing a number of dots. The subjects then communicated with each other and settled on an estimate of the number of dots. Influence, measured by the amount that a subject changed his initial estimate during the experimental session, is found to be negatively related to the centrality of the position only for the Y network. He finds, however, a positive relation between the centrality of a position and the number of leader nominations.

Trow (1957) develops a point made by Leavitt (1951, p. 49) into the hypothesis that centrality produces high morale and status not just because centrality implies greater access to communication channels, but because greater access to channels gives autonomy—ability to make independent decisions. Trow argues that though centrality and autonomy are usually correlated, they can be separated experimentally and that when they are separated, autonomy will be found to be the effective variable. He accomplished this separation by placing his subjects in apparent three-person chains and passing prepared notes to them to create the illusion of a group. Trow varied autonomy by giving some subjects a code book needed in planning the group's task and informing other subjects that someone else in the group had the code. He also gave the subjects a questionnaire to measure their need for autonomy.

The major findings are the following: autonomy produces a higher level of job satisfaction than does dependence; the effect of centrality

TABLE 3

Synopsis of Initial and Follow-up Studies

Investigator	Task	Network	Independent variable	Dependent variable	Findings[a]
Bavelas (Smith) (1950)	Determining common symbol	Chain, Circle (5-man)	Network Position Centrality	accuracy leader nomination	N→a: Ch>Cc PC→ln: +
Leavitt (1951)	Determining common symbol	Chain, Circle, Wheel, Y (5-man)	Network Position Centrality	speed accuracy leader nomination morale number of messages	N→s: 0[b] N→a: Y>Cc N→nm: Ch, Wl, Y<Cc PC→ln: + PC→mr: + PC→nm: +
Heise and Miller (1951)	Reconstruction of word lists, sentences, anagrams	Chain, one-way and two-way channel Circles (3-man)	Network Noise Task Network×Noise ×Task Interaction	speed accuracy number of words	N→s: + N→nw: + Ns→s: − Ns→a: − Ns→nw: + N×Ns×T→s, a, nw: +
Guetzkow and Simon (1955)	Determining common symbol	All-Channel, Circle, Wheel (5-man)	Network	speed organizational stability message content	N→s: Wl>A-Cl>Cc N→o st: Wl>Cc>A-Cl
Guetzkow and Dill (1957)	Determining common symbol	All-Channel, Circle (5-man)	Network Communication Restriction during Intertrial Organizing Period: Circle vs. Circle—All-Channel	speed message content	ComR→s: 0
Goldberg (1955)	Group decision on number of dots	Chain, Wheel, Y (5-man)	Network Position Centrality	"influenceability" leader nomination	PC→infl: 0 PC→ln: +
Trow (1957)	Modified common symbol problem	simulated Chain (3-man)	Position Autonomy Position Centrality Need for Autonomy	morale status	PA→m: + PA→st: 0 PC→m: 0 PC→st: +

[a] The abbreviations in the Findings column of this and the following synoptic tables are derived from the independent variable and the dependent variable. They read as follows:
N→s: + = Network (independent variable) has an effect on speed (dependent variable).
PA→m: 0 = Position Autonomy does not have an effect on morale.
If the independent variable is at least an ordinal measure, then the symbol + takes on added meaning, signifying the direction of the relationship. In these cases:
Ns→nw: + = Noise level is positively related to the number of words transmitted.
Ns→s: − = Noise level is negatively related to speed.
If the independent variable is a nominal measure, then the findings are abbreviated as follows:
N→s: Wl>A-Cl>Cc = Network affects speed, with Wheel faster than All-Channel which is faster than Circle.
Inequalities in such findings are always given with the superior groups on the left. Thus,
N→a: Y>Cc = Network affects accuracy, with Y better than Circle,
but N→nm: Wl, Y, Ch<Cc = Network affects number of messages, with Wheel, Y, and Chain better (requiring fewer messages) than Circle.

[b] The interpretation of the finding does not agree with the investigator's.

upon satisfaction is not significant. The relation holds primarily for the high-need subjects. Trow concludes that "autonomy may be considered as mediating the observed relationship [found by Leavitt] between centrality and satisfaction" (p. 208). Predictions concerning a parallel effect of autonomy on self-ascribed status were not supported. Status was, however, affected by centrality.

The studies summarized in this section exemplify the major developments of the original theme: addition of new variables, e.g., noise, and analysis of the structural variables into psychological components. A synopsis of these studies and the initial network studies is presented in Table 3.

Testing the Limits of the Basic Design

Shaw has systematically worked the area opened up by Bavelas, extending the investigations to include such variables as amount and distribution of information, problem

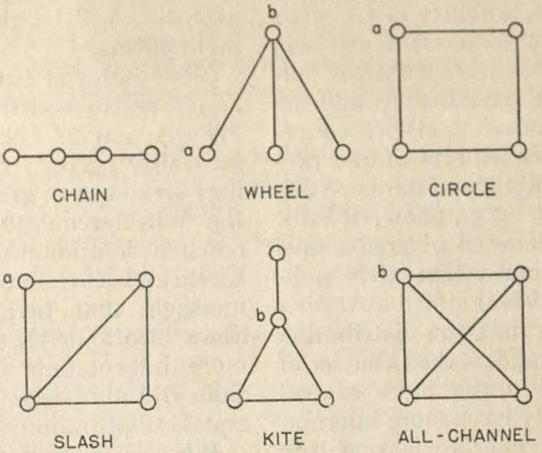


FIG. 3. Four-man networks used by Shaw.

complexity, and type of leadership. He has also suggested additional concepts: independence of positions (rather than centrality) and saturation.

Shaw (1954a) extended the network investigation to four-man groups (see Figure 3). The names he assigns to his networks raise an interesting question. Could not the four-man "Wheel" also be called a "Y"? The question has importance in comparing results for networks that differ in size. There is no empirical or rational basis for matching results from a four-man and five-man "Wheel." The only thing clear is that the number of distinct patterns decreases as the number of group members decreases. Therefore, although Chain, Wheel, and Y are distinct patterns for five-man groups, when the number of members is reduced by eliminating a peripheral member, only two of these three patterns remain: four-man Chain and Wheel-Y. If the number of members is reduced again, the two remaining networks coalesce into the simple three-man Chain. The difficulty caused by ignoring this characteristic will be pointed out later.

Shaw finds that the centrality of position is related, as in the Leavitt study, to number of messages sent, satisfaction, and frequency of nomination as leader. He proposes, however, an alternative to centrality, the related concept of independence, I, and constructs a measure of it. Shaw then plots mean number of messages, morale, recognition of leadership against I for his own and Leavitt's data. I appears to give plots that are more nearly monotonic than does centrality. The functions are, however, not only complicated, but also differ in form for presumably comparable Shaw and Levitt data. For example, the equation relating morale to I is logarithmic for Leavitt's data and linear for Shaw's. The need for a concept like independence in explaining behavior within the networks had been expressed by Leavitt (1951) and has received empirical support by Trow (1957). Shaw's I, however, is an awkward and complex combination of variables. Since he gives no statistical evaluation of the improvement of I's fit of the data over the fit yielded by centrality, it is difficult to judge whether I's smoother curves compensate for its greater complexity. (The comparability of Leavitt's and of Shaw's data is of concern. Shaw told his subjects what the network structure was. Leavitt and the investigators following his procedures did not. Other aspects of the presumed comparability of data are discussed later on—e.g., Shaw, 1954b).

The data Shaw considered above were drawn from a separately published study (1954c) aimed at testing the hypothesis that the distribution of information affects the behavior of networks. Since the more central positions usually have more information than the other team members during the major part of a trial, the effects of centrality and amount of information are ordinarily confounded. Shaw, therefore, varies the amounts of information initially given to positions within three networks. In this way, he separates to some extent the effects of the two variables.

He uses the three four-man communication patterns depicted in Figure 3: Circle, Wheel (or Y), and Slash. The groups solved arithmetic problems for which each team member held some of the necessary information. In some teams, all members had the same amount of information. In other teams, the information was unequally distributed with the positions marked *a* in Figure 3 receiving more information than the others.

He finds that central positions and the positions with the larger amount of initial information tended to solve the problems more quickly. There were no significant effects of networks or distribution of information conditions on network speed. Here, and in the following studies, Shaw centers much of his data analysis on the higher order interactions, e.g., network with information distribution with trials. Since his hypotheses and his conclusions are not at this level, attention will be given primarily to main effects.

The results on number of messages as related to network (Circle versus Wheel) and as related to position centrality agree with Leavitt's findings for five-man groups. In general, the Wheel required fewer items to reach a solution than the Slash or Circle and central positions sent more messages than peripheral positions. Shaw also finds that positions given more information sent more items than did the same positions under equal distribution of information.

What is the meaning of a relation between the number of messages—a measure used by Leavitt, Shaw, and the investigators who follow them—and position differences? Since the different positions have to send different minimum numbers of messages to complete a trial, it is not very enlightening to note that differences appear. In a five-man chain with each man holding one item of information, the end men have to send only one message in order to assure complete distribution of their information. The central man has to transmit five items. It is necessary to relate the number of messages sent by a position to the minimum for the position. Otherwise, it is as if an experimenter reported significant differences in the number of responses by two experimental groups when one group of subjects is requested to name two items each, the other requested to name only one.

Shaw does not find that differences in network affect the number of errors, although unequal distribution of information lowers the number of errors significantly. On the other hand, leadership, measured in terms of preference in a sociometric questionnaire, was related to centrality but not to information distribution.

Similarly, group morale measures and individual morale or job satisfaction ratings were, as in the Leavitt study, related to centrality. They were not, however, related to information distribution.

Gilchrist, Shaw, and Walker (1954) explored the effect of distribution of information further by giving additional information not only to peripheral, but also to central positions in the four-man wheel. Their three experimental conditions consisted of an equal distribution of information, an unequal distribution to the periphery (one peripheral subject receiving more information than the others), and an unequal distribution to the center (the center subject receiving more information). Distribution of information did not have any significant effect on overall group performance as seen in time scores, error scores, sociometric choices, number of message units, and leadership emergence. It did have an effect on behavior at individual positions. Increasing the initial information, in general, decreased the time scores and increased the number of messages transmitted, job satisfaction, and position status rating. The investigators' expectations concerning the order of the time scores are not met. The central position with additional information has a higher time score than the peripheral positions with additional information and also a higher time score than the central position under equal information distribution. Primarily to explain the latter result, they introduce the concept of saturation, defined as the input and output requirements which are imposed on positions within a group structure. The concept, a promising one which suggests that communication requirements may counteract the effects of centrality, is explored in a subsequent investigation (Shaw, 1955).

Shaw (1956) investigated the effect of another aspect of information distribution in communication networks: random versus systematic distribution. In solving an arithmetic problem consisting of four distinct steps, each member of a four-man group may have all the information items necessary to complete one of the steps. This is called systematic distribution. A random distribution is one in which each of the information items is assigned at random; a member, therefore, usually has to go to several sources (other members) for the information to complete one step of the problem. This type of experimental operation brings the network study close to the situations used by Lanzetta and Roby in their manipulation of "dispersion of information sources" (see below). Shaw predicts that systematic distribution will increase efficiency and job satisfaction and that the increase will be greater if the subjects are informed about the system of distribution and if the network permits freedom of action—e.g., All-Channel as compared with Wheel. Analysis of time to solution in the Wheel and All-Channel networks, in part, supports Shaw's predictions. Knowledge of distribution is not, however, significant as a main effect or in interaction with distribution of information. Networks, also, do not differ significantly on the time measures.

Another follow-up (Shaw, 1954b) of the distribution of information study by Shaw (1954c) attempts to reconcile an apparent discrepancy between his and Leavitt's (1951) study. Leavitt found some evidence that the five-man Wheel network solves problems faster than the five-man Circle. Shaw's four-man Circles are somewhat faster than his four-man Wheels. The speed difference is, however, not statistically significant. Shaw suggests that the difference stems from the difference in the complexity of the problems used. This is essentially Heise and Miller's (1951) point that different structures will be best for different tasks. Shaw's (1954b) main hypothesis is

> that a communication net in which all *S*s are in equal positions (the circle) will require less time to solve relatively complex problems but more time to solve relatively simple problems than will a communication net in which one *S* is placed in a central position (the wheel) (p. 211).

To test this hypothesis, Shaw gave simple (common letter) and complex (arithmetic) problems to the three-man Wheel and Circle (Networks 3 and 1, respectively, in Figure 2).

Two points should be noted concerning his structures: they are three-man groups not five-man groups, as in the Leavitt study, and not four-man groups, as in the other Shaw study; the question raised earlier concerning the naming of networks may be raised again. Why is the third pattern in Figure 2 called a "Wheel" rather than a "Chain"? With these points in mind, it seems unlikely that differences in the results of a study of five-man groups and a study of four-man groups can be resolved by a study of three-man groups. Resolution is especially unlikely since the Chain, which, according to Leavitt, tends to be slower than the Circle, and the Wheel, which tends to be faster than the Circle, reduce to a single network in the three-man group. Shaw identifies this network with the fast Wheel. It could just as well be identified with the slow Chain. In any case, Shaw's main hypothesis is that the interaction of problem complexity with network has an effect on solution time. The results tend to support his prediction, but are not statistically significant. Analysis of the number of items communicated and errors does not add any support to the hypothesis.

The problem complexity with net centrality interaction is pursued in one more study in which Shaw (1958). manipulates complexity by the addition of irrelevant information to arithmetic problems given to the four-man All-Channel and Wheel. The evidence is again unclear. A significant effect of the interaction is found on the number of messages but not on time to solution.

Shaw (1955) has better luck with a study of the effect of saturation and independence. He elaborates these concepts and through them arrives at the variables of the classic Lewin, Lippitt, and White study (1939): autocratic versus democratic leadership. He assumes that the leader's style affects both saturation and independence: "autocratic" leaders decrease both the independence and saturation of the followers and "democratic" leaders increase both. Independence is assumed to improve performance and morale, with a greater effect on morale. Saturation is assumed to lower performance and morale, with a greater effect on performance. From these assumptions, Shaw derives the following predictions: autocratic leaders will promote better performance than democratic leaders, autocratic leaders will cause poorer morale, and differences between central and peripheral positions will be accentuated by autocratic leadership.

The two leadership conditions were used with the four-man Wheel, Kite, and All-Channel (see Figure 3) solving arithmetic problems. The subject at Position *b* in the network was assigned the role of leader and was instructed to be either "autocratic" or "democratic" in his handling of directions and suggestions. As Shaw predicted, the autocratic groups are higher in efficiency and lower in morale. Although analysis of data for individual positions confirms previous findings that the central positions solve the problems more quickly, send more messages, and have higher morale, it does not confirm Shaw's prediction that autocratic leadership will increase the difference on these measures between central and peripheral positions. It might be added that in the Lewin, Lippitt, and White study, autocratic and democratic leadership styles generated an analog of the Wheel and the All-Channel, respectively. Under autocratic leadership, for example, most of the group's communications were directed at the leader. Shaw's study, therefore, may be viewed as involving two types of manipulation of communication structure: direct manipulation through the elimination of channels and indirect manipulation through the effects of the leader's style.

In the preceding studies by Shaw and his associates, the groups worked two to four problems. The effect of prolonged experience was investigated by Shaw and Rothschild (1956). Groups in the Wheel, Slash, and All-Channel structures (see Figure 3) solved two arithmetic problems a day for 10 days. The usual analyses are made of time scores, number of message units transmitted, and satisfaction ratings. The results, to some extent, agree with the results of previous studies (Shaw, 1954c, 1955).

The merging, seen in the study on leadership style, of Shaw's network investigations with the more conventional social psychological tradition continues in a study by Shaw, Rothschild, and Strickland (1957) in which they use unstructured group discussion tasks. Each member of the group starts with all the information required for a decision. The group members have to interact only to reach an agreement on the solution. The networks differ significantly in the time required to reach a decision. The Wheel requires the longest time and the All-Channel, the shortest. The finding agrees, to some extent, with Shaw and Rothschild's findings on the same networks solving arithmetic problems. Two other experiments reported in this article investigate the effect of the position within a network upon the ability of an individual to maintain nonconforming opinion. These experiments are similar to Goldberg's experiment (1955). The results, in general, indicate that the amount of change that a subject is willing to make is a function of the amount of support and opposition he faces rather than any position characteristic. The data on the relation between position centrality and tendency to be influenced do not, however, permit clear interpretation. Goldberg, it may be recalled, finds no overall relation between centrality and tendency to be influenced.

In summary, Shaw and his associates have exhaustively worked the area opened by Bavelas and Leavitt. They have also introduced new concepts, e.g., independence and saturation, which are worth further examination. Their work forms a major body of data concerning the effect of structure on group behavior. An overall summary of these findings is presented in Table 4. A glance at the variables employed in successive studies indicates that the area has been worked not only exhaustively, but to exhaustion.

After a promising start, the approach has led to many conflicting results that resist any neat order. Perhaps more significant as a symp-

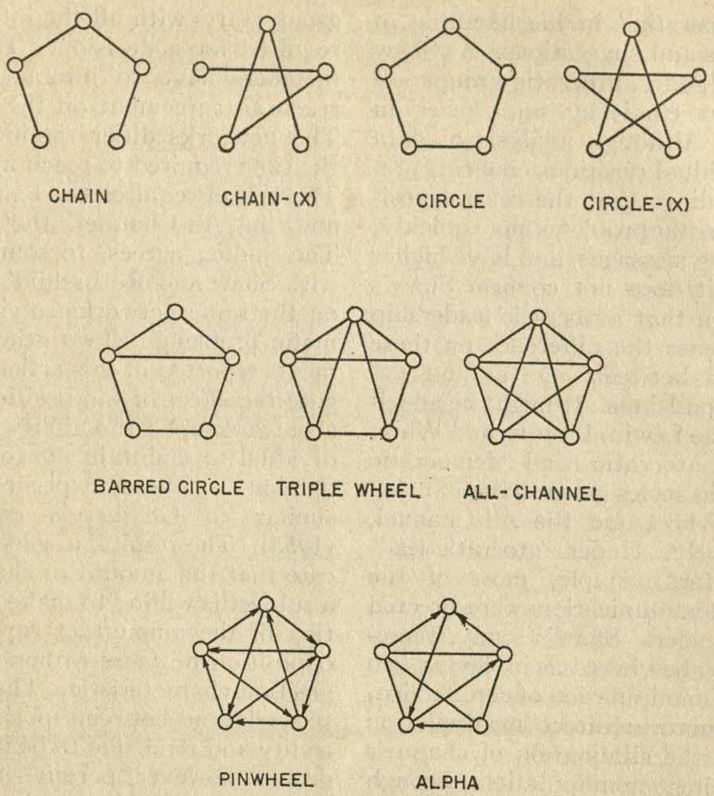


FIG. 4. Five-man networks used by Christie, Luce, and Macy. Arrows indicate one-way Channels. (The "Triple Wheel" is called "Wheel" by Christie et al.)

tom of morbidity is the lack of new hypotheses. The lack is seen in the regression to nonstructural independent variables: leadership style, conformity pressure.

Mathematical Analysis

Christie, Luce, and Macy and their associates have carried out an intensive program of investigation of behavior in group networks. They have emphasized "pure" structural characteristics and have subjected their data to detailed mathematical and logical analysis. The full range of their approaches to network behavior is set forth in two reports (Christie et al., 1952; Luce, Macy, Christie, & Hay, 1953). In the first report, they discuss the various aspects of network behavior extensively, and analyze data obtained in a series of experimental studies.

One of the studies was concerned with the effects of learning on performance in the networks in Figure 4. Christie (1954) later published results for the five-man Circle, Chain, All-Channel, and Pinwheel.[3] The groups solved a series of 25 list reconstitution problems like those in the Heise-Miller (1951) study. An "action quantization" restriction was imposed in order to simplify the data for analysis:

The subjects were required to send single-

[3] This section draws both from results presented in the larger report (Christie et al., 1952) and from the separate report by Christie (1954).

TABLE 4

Synopsis of Shaw Studies

Investigator	Task	Network	Independent variable	Dependent variable	Findings
Shaw (1954c)	Arithmetic problems	Circle, Slash, Wheel (4-man)	Distribution of Information: Equal vs. Unequal Network Position Centrality Position Information (High vs. Low Information at a Given Position)	speed accuracy number of messages leader nomination morale	DI→s: 0 DI→a: Unq>Eq DI→nm: 0 DI→ln: 0 DI→mr: 0 N→s: 0 N→a: 0 N→nm: Wl<Cc<Sl N→mr: Cc, Sl>Wl PC→s: + PC→nm: + PC→ln: + PC→mr: + PI→s: + PI→nm: + PI→ln: 0 PI→mr: 0
Gilchrist, Shaw, and Walker (1954)	Arithmetic problems	Wheel (4-man)	Distribution of Information: Equal vs. Unequal Peripheral vs. Unequal Central Position Centrality Position Information	speed accuracy number of messages leader nomination morale	DI→s: 0 DI→a: 0 DI→nm: 0 PC→s: + PC→a: 0 PC→nm: + PC→ln: + PC→mr: + PI→s: + PI→ln: 0 PI→mr: +
Shaw (1956)	Arithmetic problems	All-Channel, Wheel (4-man)	Distribution of Information: Systematic vs. Random Knowledge of Information Distribution Network Problem Difficulty	speed accuracy number of messages morale	DI→s: Sys>Rdm DI→nm: 0 DI→mr: Sys>Rdm KID→s: 0 KID→nm: 0 KID→mr: 0 N→s: 0 N→nm: Wl<A-Cl N→mr: A-Cl>Wl
Shaw (1954b)	Common letter (simple) and arithmetic (complex) problems	Circle, Wheel (3-man)	Problem Complexity ×Network Interaction Network Problem Complexity	speed accuracy number of messages	Comp×N→s: 0 Comp×N→a: 0 Comp×N→nm: 0
Shaw (1958)	Arithmetic problems	All-Channel, Wheel (4-man)	Network Irrelevant Problem Information Networks×Irrelevant Information Interaction	speed number of messages morale	N→s: A-Cl>Wl N→nm: Wl<A-Cl N→mr: A-Cl>Wl II→s: − II→nm: 0 II→mr: 0 N×II→s: 0 N×II→nm: + N×II→mr: 0
Shaw (1955)	Arithmetic problems	All-Channel, Kite, Wheel (4-man)	Leadership Style: Autocratic vs. Democratic Network Position Centrality Position Centrality ×Leadership Style Interaction	speed accuracy number of messages morale	LS→s: Auto>Demo LS→a: Auto>Demo LS→nm: Auto<Demo LS→mr: Demo>Auto N→s: 0 N→a: 0 N→nm: Wl<Kt<A-Cl N→mr: A-Cl>Kt>Wl PC×LS→s: 0 PC×LS→nm: 0 PC×LS→mr: 0
Shaw and Rothschild (1956)	Arithmetic problems	All-Channel, Slash, Wheel (4-man)	Trials: Learning Network Position Centrality	speed number of messages morale organizational structure	T→s: + T→nm: − T→mr: + N→s: A-Cl>Wl, Sl N→nm: Wl<A-Cl<Sl N→mr: 0
Shaw, Rothschild, and Strickland (1957)	Group decisions about "human relations" problems	All-Channel, Slash, Wheel (4-man)	Network Position Centrality	speed number of messages morale	N→s: A-Cl>Sl>Wl N→nm: 0 N→mr: 0 PC→s: 0 PC→nm: + PC→mr: 0
	Estimation of number of clicks	Wheel (4-man)	Position Centrality Support vs. Opposition from Other Members	"influenceability"	PC→infl: 0

address messages at prescribed times, to include in their messages all problem information known to them at the time, and to write nothing other than problem information. . . . Thus, each message-sending action, hereinafter called an *act*, was a simultaneous sending by the whole group (p. 189).

In their data analyses, these investigators use the minimum number of acts in which it is possible for a network to complete its task as baselines. The minimum possible number of acts is an important consideration which has been neglected in other network studies. All networks in Figure 4 have minima of three acts except for the Chain with a minimum of five. Chain-(X) and Circle-(X) are topologically the same as Chain and Circle, respectively. They differ from Chain and Circle only in the physical arrangement of the positions. The investigators computed the distribution of the number of acts required for completion on the assumption that the group members distribute their information at random over their channels. It is not surprising that comparison of the theoretical with the observed number of solutions show that the subjects do better than chance from the start. Clear differences in learning efficiency between networks are also demonstrated.

Christie (1954) summarizes the results for four of the networks as follows:

> Groups using the totally connected network [All-Channel] do somewhat better than random but show a negligible amount of learning. The groups in chain learn well, but their performance is good only with respect to the chain minimum of five acts per trial. The high minimum for this network makes its absolute performance poor in comparison to each of the other networks. The pinwheel network performs somewhat better than random, and its random distribution is a favorable one [i.e., the mode is close to the minimum]. Like totally connected it learns little, so that its final distribution is practically the same as that in totally connected. Circle is the one very different case; it achieves the best distribution [i.e., it frequently completes the task in the minimum possible] in the final block as a result of excellent learning (p. 193).

Christie, Luce, and Macy introduce the concept of "locally rational" behavior to explain the differences between networks on the basis of behavior at the individual positions. Locally rational behavior is the tendency to send successive messages to different stations so as to maximize the amount of new information received by neighboring positions. ". . . the behavior called for depends only on each subject's attending to conditions immediate to his own position, i.e., to whom he has sent and from whom he has received" (Christie, 1954, p. 195). The investigators used Monte Carlo runs on a computer in order to obtain the theoretical distribution of the number of acts to completion under both the equiprobable random behavior and the locally rational behavior model for each network. In a final summary of their work, Christie, Luce, and Macy (1956) show that in successive trials, network performance (the distribution of number of acts) approaches more and more closely that of the locally rational model. It is not clear, however, whether this generalization holds for the All-Channel network.

In analyzing the learning of subjects within the various networks, they also pay attention to the importance of differences in the probability of various initial act patterns. For example, in the Circle a minimum solution is possible only with an initial pattern of two mutual interchanges and one unreciprocated message (e.g., *a* with *b*, *c* with *d*, and *e* to *a*). In the Chain, any initial pattern of acts can result in a minimal solution.

They find another stimulus for

experimental work and analysis in information theory concepts. In a study on coding noise presented in the main technical report (Christie et al., 1952) and published separately by Macy, Christie, and Luce (1953), they examine the effects of ambiguity of stimuli, interpreted as semantic noise. (Heise and Miller, 1951, studied the effect of acoustic noise in the communication channel.) They used the five-man Circle, Chain, Wheel, and Pinwheel (see Figures 1 and 4). Another variable, labeled "feedback," is introduced by giving some Wheel groups additional information concerning errors at the end of each trial.

The groups' task was to discover the color of a marble that all held in common. Fifteen problems with marbles of clearly identifiable color were followed by 15 trials with ambiguous stimuli—marbles of mixed, indistinct color. The authors state that the data on speed and number of messages agree with Leavitt's results. Their main findings on learning to handle the ambiguous stimuli do not give a simple picture. The Circle reduces its errors markedly over successive trials. The other structures do not. The explanation of these results may lie in Shaw's (1954b) hypothesis, as yet undemonstrated, that centralized structures are handicapped on complex problems. Introduction of additional feedback in the wheel network seems to improve performance somewhat.

They pursue their informational analysis with an estimate of "conditional receiver entropy" based on the number of different marbles called by the same name. They point out that the method by which the efficient networks reduce ambiguity seems to be by an increase in redundancy (computed in terms of the number of extra names given to a marble). The behavior of the networks is further analyzed qualitatively in terms of error feedback (the opportunity of the members to obtain the same information from at least two different sources) and the opportunity to correct errors (the presence of symmetric, i.e., two-way channels). The Pinwheel lacks the latter, while the Wheel, and to some extent the Chain (in its end members), lacks the former. The presence of both, the investigators argue, is necessary for optimal performance.

Christie et al. (1952) try to carry out detailed analysis and derivation of every type of data generated by the original network studies. They try to derive the distribution of group latency data on the basis of assumptions concerning the individual latency distribution. They analyze the determinants of leader designation, using an index based on

> the relative frequency of use of a channel (on an equiprobable sending basis) and the mean input of the sending end of the channel as an estimate of the sending end's value to the receiving end (p. 179).

Their index fits the obtained values for two networks rather well.

They devote a similarly detailed analysis to the determinants of job satisfaction, following their general approach of using the individual position as a basis for prediction. An index called input potential which considers the input density for each position is found to be more highly correlated with job satisfaction than peripherality. The formula for input potential gives some idea of the level of the analysis.

$$g(I) = \frac{1}{\pi} \sin^{-1} \left\{ 2\frac{e^{2I}-1}{e^{2I}+1} - 1 \right\} + \frac{1}{2} \quad [2]$$

where I is mean input.

In a subsequent report (Luce et al., 1953), they carry out the same types

TABLE 5

Synopsis of Christie, Luce, Macy Studies

Investigator	Task	Network	Independent variable	Dependent variable	Findings
Christie (1954) also in Christie, Luce, and Macy (1952)	Reconstruction of number list	All-Channel, Chain, Circle, Pinwheel (5-man)	Network Trials: Learning	amount of learning number acts to solution	N→al: Cc, Ch>Pw, A-Cl
Macy, Christie, and Luce (1953) also in Christie, Luce, and Macy (1952)	Determining common ambiguous marble	Chain, Circle, Pinwheel, Wheel (5-man)	Network Additional Feedback in Wheel	accuracy	N→a: Cc>Wl, Ch, Pw

of analyses and also examine the additional problem of the effects of change of network structure, with subjects trained in one network and tested in another. The later work is even more complex, involving a multiplicity of approaches, and does not lend itself readily to summary. Since their work is primarily concerned with analytic techniques, Table 5, which includes only those empirical results that are comparable to the results in the other summary tables, does not do justice to the full range of their work. The general philosophy and major accomplishments of their research effort is summarized by Christie et al. (1956).

The main tendency of the analysis by Christie, Luce, and Macy is away from functions involving overall measures of the group, e.g., network centrality, and toward the derivation of the behavior of the group from that of the individual positions. Their efforts may be considered to parallel Shaw's. Just as he carried the empirical work in the area as far as it can go, so do they carry the mathematical analysis to the limit. In both cases, it was desirable to have the job done. It seems unlikely now, however, that the payoff will be commensurate with the energy and ingenuity that was invested. This could, of course, be only discovered by the doing.

With the efforts of both Shaw and his associates and of Christie, Luce, and Macy carried as far as they can go, a new approach or new definition of the field seems necessary. The remaining sections of this paper will review some of the attempts at reanalyzing or redefining the area.

Retrospect and Prospect

At this point, it is appropriate to look back at the problem as originally stated and its expression in experimental form.

Two main questions posed by Bavelas are the following: What effect does the structure of the group have upon its efficiency? What effect does position in the group have upon the subject's morale and job satisfaction? There is no simple answer to the first question. The effect of structure depends in part on the requirements of the task (Heise & Miller, 1951). Contrary to Leavitt's original generalization (1951), in a number of studies the highly centralized structures are *less efficient* than other structures (Macy et al., 1953; Shaw, 1958; Shaw et al., 1957). The answer to the second question is somewhat clearer. Morale seems to be a function of centrality of position. The psychological basis for this relationship, however, warrants further analysis. Explanations have been offered in terms of autonomy (Trow, 1957), independence (Shaw, 1954a), and input potential (Christie et al., 1952).

The unclear answers to the first question may arise from the peculiar experimental situation used to express it. The characteristics of the original Bavelas-Leavitt situation that recommend it are its apparent experimental simplicity and relevance to real-life situations. Does the situation actually have these characteristics? That the situation is not simple is evidenced by the introduction of techniques to simplify it further, e.g., action quantization. Even with the imposition of further restrictions, however, a precise analysis of the activity of the groups is unmanageably complex.

That the situation is far distant from most familiar real-life situations can be seen by reviewing the special characteristics of the laboratory networks. They are the following:

1. Interdiction of certain channels. This is the most obvious of the special characteristics of the laboratory networks. To some extent, this corresponds with conditions in natural groups. Some communication channels are frequently closed to members of groups. For example, a man may not be permitted to go over the head of his immediate supervisors in a work group, or he may be unwilling to make certain statements when another member is present.

2. Ignorance concerning other positions. This is probably both the effective and the really unique aspect of the communication restriction. The network members know very little about other positions and about behavior of any except adjacent positions. This is a condition that does not hold in small groups. The effect of this factor can, of course, be reduced to some extent by changes in the procedures of the network studies. In the Guetzkow-Simon (1955) study, for example, this may have been done by having intertrial administrative discussions.

3. Necessity of each member. In almost all the network studies, each member is essential, because each member holds an essential piece of information and each member must present a solution to the problem. In some cases, one member may have more or less information but in almost all the studies the elimination of one member prevents success of the group. This is not generally true in real-life groups.

These special characteristics of the network studies would make generalization difficult even if the findings were unequivocal. The applicability of the findings of the network studies are in question because the characteristics of the structures employed in the studies are very different from other small groups. The following point, however, may be argued: If the network studies have any application, it will not be in the small group, but in a much larger unit such as an industrial corporation or an army. Characteristics analogous to those listed above are more clearly present in large groups. For example, departments of a company may not have direct communication channels; they often lack information concerning distant sections and all departments may be necessary for the company to function.

If the laboratory network cannot be viewed as a simplification of the general small group situation, can it be viewed as a laboratory simplification to permit testing of an explicit theory about group behavior? The answer, unfortunately, is no. At the present time, a theory concerning behavior in the network does not exist. This raises a major point. Perhaps the most surprising thing about the entire area is that despite the highly formal origins of these studies (Bavelas, 1948), the organized body of theory promised by the approach has not yet appeared.

Perhaps in response to considerations such as these, two attempts have been made to use a somewhat different approach to the study of the effects of group structure on behavior. One of these attempts is by Lanzetta and Roby. Their main aim is to draw the experimental situation closer to a known type of group—the military work team. The other attempt is by Rosenberg and Hall. Their main objective is to simplify the experimental situation (two-person situations) and rephrase the problem so that available theory—learning theory—can be brought to bear on the problem. Both approaches assign new definitions to the term structure. For Lanzetta and Roby, team structure refers to the specialization and interrelation of jobs in a team. For Rosenberg and Hall, team structure refers to the degree to which the information that an individual receives about his performance is confounded by the performance of another team member.

Emphasis on Distribution of Functions in the Simulated Team

Lanzetta and Roby have directed one major attempt to examine, from a new viewpoint, the relation of group structure to performance. Their attempt is embodied in a series of studies in which they vary the ways that team members depend on each other for information. In a situation, quite unlike the Bavelas network, modeled after military teams, e.g., a bomber crew, they gave teams a series of very short problems in order to approximate a continuously changing environment. They vary communication structure not by interdicting channels as in the network studies, but by restricting relevant information or specific functions to a given position. Their team is like the All-Channel network but with each subject working on a separate problem and holding some information required by other team members. Despite these differences in experimental situation and in definition of structure, the basic concern remains the same: what factors in the organization of a group affect its performance?

An early study by Lanzetta and Roby (1956b) indicates both the development of their experimental situation and the type of practical situation from which it grew. In this study, they investigate the effect of two methods of work distribution (work structure) under two task load conditions on group performance. They model the experimental situation after an air defense center with two work structure conditions. In vertical structure, each group member had one of three tasks: tracking aircraft, identifying aircraft and keeping a record of the interceptors' fuel status, or deploying friendly planes. In horizontal structure, each group member performed all three functions for his own targets. Variations of the number of airplanes produced two different task load conditions. Of the main independent variables of the study—structural organization, load conditions, and their interaction—only load condition has a significant effect. The interpretation of this effect is, however, complicated by a significant interaction with sessions. The main outcome of the study was a methodological development rather than an empirical finding. It led to a simpler task with higher reliability for use in the subsequent studies.

This task, modeled after a bomber crew's task, was used in their next experimental study (Roby & Lanzetta, 1956a) to demonstrate the effect of relaying requirements upon group performance. Groups of three subjects sat, each in a separate booth that contained instrument reading displays, pairs of control switches, and instructions giving the correct switch settings for each possible pair of instrument readings. The instrument readings required to set a given control could be displayed in the booth containing the related control or they could be shown in one of the other booths. In the latter case, the subject receiving the information would have to relay it to its eventual user over the intercom system connecting the booths.

Roby and Lanzetta used four communication structures which differed in the degree to which the subjects had direct access to the information they needed. A significant difference in the number of errors appears between the communication conditions. Analysis indicates that more errors are made on a control if its two relevant items of information had to come from two sources rather than from one source. The results cannot be considered surprising. If a subject has to get the necessary information from someone else, who is also busy, he will not do as well on a highly speeded task as a subject who has his information immediately available. Furthermore, if he has to make two separate information requests in a brief (15-second) period, he will be more likely to fail than if he has to make only one request.

Lanzetta and Roby (1956a) next consider the effect of type of input presentation on efficiency in two communication structures employed in the previous study: a high dependence (or low autonomy) condition in which each member had to get all of the instrument readings necessary to operate his controls from other team members; and a low dependence condition in which a member had three out of the four necessary instrument readings available in his booth. They varied two aspects of the task input: task load (the time interval between successive presentations of instrument readings) and the predictability of the order of presentation to the three booths. They find again that high dependence gives rise to more errors, especially when the information has to be relayed from several different sources. For both structure conditions, errors increase as the rate of change of instrument readings increase, but predictability of the order of instrument changes has no significant effect.

These findings are further supported in another study in which Lanzetta and Roby (1957) investigate learning and the details of communication behavior in their team situation. They vary dependence (relaying requirement), task load (speed of presentation of input), and operating procedure as determined by instructions to "volunteer" information or to "solicit" information.

In a later study, Roby and Lanzetta (1957) consider the effect of "load balancing" or distribution of work. They used three structures that varied the relation between the number of instrument displays and the number of control switches for which the subject was responsible. In Structure I (equal observation load) a booth had either one, two, or three control switches, but it always had two displays. In Structure II (unequal load) a booth that had one control switch had one display; a booth with two control switches had two displays; etc. In Structure III (balanced load) a booth with one control switch had three displays; a booth with two control switches had two displays; etc. The experimental design is quite complicated and confounds the load balancing and dependence variables. The authors, however, conclude that "both load balancing and autonomy are influential but that the latter is more heavily weighted in this task" (p. 174).

The major accomplishment of Lanzetta and Roby is their introduction of controlled, experimentally manipulable tasks that capture more of the characteristics of real-life teams than do the earlier Circles and Wheels. They have also theorized extensively (Roby, 1957; Roby & Lanzetta, 1956b, 1958). The real payoff in their work will come, however, when theory and experimental work merge. Their theorizing consists of general statements that never arrive at the prediction or explanation of specific events. Without a theory to generate novel and testable predictions, the experiments usually establish the obvious, e.g., if a subject has to check with many people before he makes a response, he is not likely to complete the response in a short time period. Although Lanzetta and Roby have not completed the merger of theory and experiment, they have brought them several steps closer together. A summary of their experimental findings is presented in Table 6.

Emphasis on Feedback and Learning

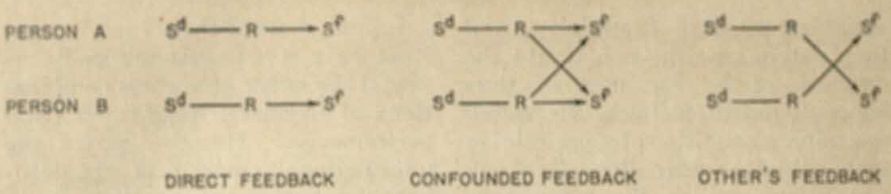
Rosenberg and Hall have recently examined the effects of group structure from a different viewpoint than Lanzetta and Roby's. Rosenberg and Hall see the composition of information feedback to the individual members as a key aspect of structure. They concern themselves, therefore, with the relation of structure, defined in terms of information feedback, to performance. Figure 5 illustrates the basic structures they study. S^d is the stimulus which precedes a response, R is the response, and S^f is the feedback stimulus, i.e., the state of affairs in which the individual finds himself after performing the response. In the "direct" feedback condition the S^f reflects only the subject's own per-

TABLE 6

Synopsis of Lanzetta and Roby Studies

Investigator	Task	Independent variable	Dependent variable	Findings
Lanzetta and Roby (1956b)	Simulated air defense center	Structure: Horizontal vs. Vertical Work Load: Number of Aircraft	weighted efficiency score	St→e: 0 WL→e: −
Roby and Lanzetta (1956a)	Simulated military crew: dial reading and switch setting	Dependence (on Others for Information) Dispersion (of Information Sources)	accuracy	Dp→a: − Ds→a: −
Lanzetta and Roby (1956a)	Simulated military crew	Dependence Dispersion Task Load: Input Presentation Rate Predictability of Input	accuracy	Dp→a: − Ds→a: − TL→a: − P→a: 0
Lanzetta and Roby (1957)	Simulated military crew	Dependence Task Load Operating Procedure: Volunteering vs. Soliciting Information Trials (Learning)	accuracy communication measures: number of messages length of messages total talking time	Dp→a: − Dp→cm: + TL→a: − TL→cm: 0[a] OP→a: 0 T→a: + T→nm: + T→lm: − T→ttt: +
Roby and Lanzetta (1957)	Simulated military crew	Dependence Load Balancing: Work Distribution Trials Task Load	accuracy number of messages (findings on nm replicate Lanzetta and Roby, 1957, above)	Dp→a: − LB→a: + T→a: + TL→a: −

[a] Corrected for length of trial.



DIRECT FEEDBACK CONFOUNDED FEEDBACK OTHER'S FEEDBACK

FIG. 5. Feedback conditions used by Rosenberg and Hall.

formance. With "confounded" feedback the response of one subject combines with that of another so that his 'feedback is a function of his teammate's performance as well as his own. With "other's" feedback the subject receives feedback solely from someone else's performance. In order to investigate the relation of these structures to performance, Rosenberg and Hall have carried out a series of studies using variations of an experimental situation similar to that of Sidowski, Wyckoff, and Tabory (1956).

In their first study, Rosenberg and Hall (1958) ran two-man groups under the three structures described above. The task was to learn to turn a knob a required number of turns. The amount of error (S^f value) was displayed to the subject after each trial. Under direct feedback, each subject had to learn to turn the knob four times. Under confounded feedback, the two team members had to attain a team average of four turns. They could reach this average by totaling eight turns distributed in any fashion between them. Under "other's" feedback, the subject had a perfect score displayed only if his partner turned the knob four times. The design of the study permitted the evaluation of both the effects of the subject's own feedback condition and his partner's feedback condition (which could be different) upon the subject's performance. The dependent variables were: individual accuracy, team or average accuracy, and role differentiation—a function of the absolute difference between the response magnitudes of the two team members.

The subjects learn most rapidly and to the highest level of proficiency under direct feedback. With confounded feedback the subject learns, but more slowly and to a lower level of proficiency. There is no improvement in individual accuracy under "other's" feedback. The partner's feedback condition has no significant effect on the subject's accuracy. With respect to team product, confounded feedback yields team accuracy (average performance) that is at least as good as that obtained with direct feedback. "Other's" feedback gave clearly inferior team performance. In the confounded feedback condition, one subject evidently learned to make two turns if his partner persisted in making six turns so that both subjects would have an average of four. Rosenberg and Hall label this compensatory difference between response magnitudes, role differentiation. They find that the confounded feedback conditions shows more role differentiation than the direct feedback. The "other's" feedback condition, however, shows the greatest amount of all. Rosenberg (1959b) also considered the effect of switching subjects from a direct feedback situation to other structures. After the switch, the three structures show the same effects as above.

Hall (1957), using similar apparatus, investigated two independent variables: type of pretraining, and the relative weights assigned to the responses of the team members during confounded feedback. He varied pretraining conditions by pretraining some subjects under direct feedback and others under the same confounded feedback conditions they received during later trials. The experimenter used two confounded feedback weightings—equal and unequal. In equal weighting, he fed back the mean of the two members' responses or

$$s^f = \tfrac{1}{2} R_1 + \tfrac{1}{2} R_2$$

as in the previous experiments. In the unequal weighting, he weighted the responses of one member three times as heavily as the other, i.e.:

$$s^f = \tfrac{3}{4} R_1 + \tfrac{1}{4} R_2$$

The dependent variables were team accuracy and role differentiation. The feedback weighting conditions do not have any significant effect on the dependent variables during either pretraining or training. In discussing the results, Hall emphasizes the compensatory behavior that occurs in the confounded feedback situation.

Zink (1957) carried out a further study in this series using a more complex task and a different rule for determining feedback. Contrary at least to the reviewers' expectations, the results indicate greater role differentiation for the simple task than for the complex task. Rosenberg (1959a), later tried to produce role differentiation in Zink's complex task by pretraining the subjects under direct feedback. His hypothesis was that the subjects had not reached that level of proficiency in Zink's complex task to permit them to adjust to a partner's behavior. He is, however, unable to obtain differences in role differentation between subjects given different amounts of direct feedback pretraining.

In a final set of three experiments, Rosenberg (1960) systematically explored the effect of various combinations of feedback weights on team performance. He also varied the informational content of the feedback by letting some groups know only that an error had occurred and by informing other groups about both the occurrence and the direction of the error. On the basis of detailed consideration of the effects of feedback upon the response of the subjects in the various structures, Rosenberg makes predictions concerning the development of complementary or cooperative behavior. In general, he finds that more stable response patterns develop as the amount of information concerning the direction of errors increases. If both subjects have a feedback weight of .50 or more on their own response, then their combined responses tend to stabilize at some optimal value, i.e., one in which both members receive maximum reinforcement. The accuracy of the group product is therefore maximized.

With these experiments, this group has moved very far from the original network studies. In their earlier work (Rosenberg & Hall, 1958), communication between the team members dropped out as an explicit independent variable. In the last study, amount of reinforcement received replaces group accuracy as the dependent variable of primary interest. The work of Rosenberg and Hall has certain basic similarities to the work of Lanzetta and Roby. Here again the experimenters accomplish a very able reduction of the real-life team to laboratory proportions. The contribution with respect to methods is considerable. Here again, however, the work generates obvious results. The one study (Rosenberg, 1960) with novel and systematically related results is one that has moved away almost completely from the variables of the early studies of group structure. The work of Rosenberg and Hall is summarized in Table 7.

It is hoped that as the methods in the area are improved, theories which can tie together disparate findings and generate new predictions will develop. Rosenberg and Hall have done even more than help prepare the methodological groundwork for the phase of theorizing that is needed now. By reducing social interaction to feedback conditions, they have prepared the way for an attack with the armament of learning theory. (This has actually begun in work being carried out by Burke, 1959, and his associates.) Whether such an attack can be made without giving up the original objective of studying group structure remains to be seen.

Summary and Conclusions

Since the initial stimulus provided by Bavelas in 1948 there has been a considerable effort spent on the study of the effect of structure upon group and individual behavior. The main original questions posed were: What effect does the structure of the group have upon the efficiency of its behavior? What effect does position in the group have on morale and job satisfaction? There is no clear answer to the first question. The answer to the second question is that central positions in general are more satisfied with their tasks than peripheral positions.

Later investigators went beyond the first two questions to study other variables. Heise and Miller introduced a task complexity variable, the condition of communication interference (noise), and one-way channels. Guetzkow and his collaborators introduced the distinction between task behavior and organizational activity. Shaw continued the original trend of the experimental work and also investigated the effects of various types of distribution of information and task complexity. Christie, Luce, and Macy brought mathematics and information theory to bear on the communication networks. They presented the theory of "locally rational" behavior to explain learning in the

TABLE 7

Synopsis of Rosenberg and Hall Studies

Investigator	Task	Independent variable	Dependent variable	Findings
Rosenberg and Hall (1958)	Dial turning	Subject's Feedback Condition: Direct, Confounded, "Other's" Partner's Feedback Condition	individual accuracy team accuracy role differentiation	SF→ia: D>Cf>0th PF→ia: 0 SF→ta: D, Cf>0th SF→rd: 0th>Cf>D
Rosenberg (1959b)	Dial turning	Subject's Feedback Condition (Pretraining on Direct Feedback) Partner's Feedback Condition	individual accuracy	SF→ia: D>Cf>0th PF→ia: 0
Hall (1957)	Dial turning	Subject's Feedback Condition in Pretraining Subject's Feedback Condition in Training (Two Types of Confounding)	team accuracy role differentiation	SFP→ta: 0 SFP→rd: 0 SFT→ta: 0 SFT→rd: 0
Zink (1957) also in Rosenberg (1959a)	Display-control	Task Complexity Degree of Individual Pretraining	role differentiation	TC→rd: simple>complex DIP→rd: 0
Rosenberg (1960)	Dial turning	Subject's Feedback Condition: Direct, Various Types of Confounding Information on Direction of "Error"	amount of reinforcement	SF→ar: + I→ar: +

networks and differences in performance between networks.

Neither the straight empirical work nor the mathematically sophisticated analyses have approached the goal, implicit in Bavelas' original questions, of a rational system for arranging groups to maximize efficiency and satisfaction. The difficulties in building such a system may stem from the peculiar characteristics of the Bavelas network and the absence of a theory to order the data it generated.

In response to these difficulties, more recent investigators have reoriented the work on group structure. Lanzetta and Roby have redefined structure into terms of direct versus indirect accessibility of task information and distribution of task information. Under this type of definition they have constructed new types of groups and tasks. These investigators also have made some moves toward meeting the need for a theory in the area. Rosenberg and Hall have attempted to rephrase the problem and redesign the experimental setting so that learning theory can play the organizing role. To do this, they define structure in terms of the effect of one subject's responses on another subject's feedback (reinforcement) and have studied the effect of various feedback arrangements on group (dyad) and individual behavior.

At the present time, there is still a major need for a system to order the data already obtained and to direct further work on the effects of group structure. The difficulty in constructing this system may arise from the inappropriateness of either the experimental situations or the concepts that have been used. Attempts have been made to remedy both of these possible defects. The success of these attempts will determine whether this review is a prologue or an epitaph.

REFERENCES

Bavelas, A. A mathematical model for group structures. *Appl. Anthrop.*, 1948, **7**, 16–30.

Bavelas, A. Communication patterns in task-oriented groups. *J. Acoust. Soc. Amer.*, 1950, **22**, 725–730.

Burke, C. Applications of a linear model to two-person interactions. Paper read at Midwestern Psychological Association, Chicago, May 1959.

Christie, L. S. Organization and information handling in task groups. *J. Operat. Res. Soc. Amer.*, 1954, **2**, 188–196.

Christie, L. S., Luce, R. S., & Macy, J., Jr. Communication and learning in task-oriented groups. *MIT Res. Lab. Electronics tech. Rep.*, 1952, No. 231.

Christie, L. S., Luce, R. D., & Macy, J., Jr. Information handling in organized groups. In J. F. McCloskey & J. M. Coppinger (Eds.), *Operations research for management.* Vol. II. *Case histories, methods, information handling.* Baltimore: Johns Hopkins Press, 1956. Pp. 417–537.

Gilchrist, J. C., Shaw, M. E., & Walker, L. C. Some effects of unequal distribution of information in a wheel group structure. *J. abnorm. soc. Psychol.*, 1954, **49**, 554–556.

Glanzer, M., & Glaser, R. Techniques for the study of group structure and behavior: I. Analysis of structure. *Psychol. Bull.*, 1959, **56**, 317–332.

Goldberg, S. C. Influence and leadership as a function of group structure. *J. abnorm. soc. Psychol.*, 1955, **51**, 119–122.

Guetzkow, H., & Dill, W. R. Factors in the organizational development of task-oriented groups. *Sociometry*, 1957, **20**, 175–204.

Guetzkow, H., & Simon, H. A. The impact of certain communication nets upon organization and performance in task-oriented groups. *Mgmt. Sci.*, 1955, **1**, 233–250.

Hall, R. L. Group performance under feedback that confounds responses of group members. *Sociometry*, 1957, **20**, 297–305.

Heise, G. A., & Miller, G. A. Problem solving by small groups using various communication nets. *J. abnorm. soc. Psychol.*, 1951, **46**, 327–335.

Lanzetta, J. T., & Roby, T. B. Effects of work-group structure and certain task variables on group performance. *J. abnorm. soc. Psychol.*, 1956, **53**, 307–314. (a)

LANZETTA, J. T., & ROBY, T. B. Group performance as a function of work-distribution patterns and task load. *Sociometry*, 1956, **19**, 95–104. (b)

LANZETTA, J. T., & ROBY, T. B. Group learning and communication as a function of task and structure "demands." *J. abnorm. soc. Psychol.*, 1957, **55**, 121–131.

LEAVITT, H. J. Some effects of certain communication patterns on group performance. *J. abnorm. soc. Psychol.*, 1951, **46**, 38–50.

LEWIN, K., LIPPITT, R., & WHITE, R. K. Patterns of aggressive behavior in experimentally created "social climates." *J. soc. Psychol.*, 1939, **10**, 271–299.

LUCE, R. D., MACY, J., JR., CHRISTIE, L. S., & HAY, H. D. Information flow in task-oriented groups. *MIT Res. Lab. Electronics tech. Rep.*, 1953, No. 264.

MACY, J., JR., CHRISTIE, L. S., & LUCE, R. D. Coding noise in a task-oriented group. *J. abnorm. soc. Psychol.*, 1953, **48**, 401–409.

ROBY, T. B. On the measurement and description of groups. *Behav. Sci.*, 1957, **2**, 119–127.

ROBY, T. B., & LANZETTA, J. T. An investigation of task performance as a function of certain aspects of work-group structure. *USAF Personnel Train. Res. Cent. res. Rep.*, 1956, No. TN-56-74. (a)

ROBY, T. B., & LANZETTA, J. T. Work group structure, communication, and group performance. *Sociometry*, 1956, **19**, 105–113. (b)

ROBY, T. B., & LANZETTA, J. T. Conflicting principles in man-machine system design. *J. appl. Psychol.*, 1957, **41**, 170–178.

ROBY, T. B., & LANZETTA, J. T. Considerations in the analysis of group tasks. *Psychol. Bull.*, 1958, **55**, 88–101.

ROSENBERG, S. A laboratory approach to interpersonal aspects of team performance. *Ergonomics*, 1959, **2**, 335–348. (a)

ROSENBERG, S. The maintenance of a learned response in controlled interpersonal conditions. *Sociometry*, 1959, **22**, 124–138. (b)

ROSENBERG, S. Cooperative behavior in dyads as a function of reinforcement parameters. *J. abnorm. soc. Psychol.*, 1960, **60**, 318–333.

ROSENBERG, S., & HALL, R. L. The effects of different social feedback conditions upon performance in dyadic teams. *J. abnorm. soc. Psychol.*, 1958, **57**, 271–277.

SHAW, M. E. Group structure and the behavior of individuals in small groups. *J. Psychol.*, 1954, **38**, 139–149. (a)

SHAW, M. E. Some effects of problem complexity upon problem solution efficiency in different communication nets. *J. exp. Psychol.*, 1954, **48**, 211–217. (b)

SHAW, M. E. Some effects of unequal distribution of information upon group performance in various communication nets. *J. abnorm. soc. Psychol.*, 1954, **49**, 547–553. (c)

SHAW, M. E. A comparison of two types of leadership in various communication nets. *J. abnorm. soc. Psychol.*, 1955, **50**, 127–134.

SHAW, M. E. Random versus systematic distribution of information in communication nets. *J. Pers.*, 1956, **25**, 59–69.

SHAW, M. E. Some effects of irrelevant information upon problem-solving by small groups. *J. soc. Psychol.*, 1958, **47**, 33–37.

SHAW, M. E., & ROTHSCHILD, G. H. Some effects of prolonged experience in communication nets. *J. appl. Psychol.*, 1956, **40**, 281–286.

SHAW, M. E., ROTHSCHILD, G. H., & STRICKLAND, J. F. Decision processes in communication nets. *J. abnorm. soc. Psychol.*, 1957, **54**, 323–330.

SIDOWSKI, J. B., WYCKOFF, L. B., & TABORY, L. The influence of reinforcement and punishment in a minimal social situation. *J. abnorm. soc. Psychol.*, 1956, **52**, 115–119.

TROW, D. B. Autonomy and job satisfaction in task-oriented groups. *J. abnorm. soc. Psychol.*, 1957, **54**, 204–209.

ZINK, D. L. The development of role differentiation in dyads as a function of task complexity. *Amer. Psychologist*, 1957, **12**, 371. (Abstract)

(Received October 7, 1959)

Psychological Bulletin
1961, Vol. 58, No. 1, 28–45

ON SECONDARY REINFORCEMENT AND SHOCK TERMINATION[1]

ROBERT C. BECK[2]
University of Illinois

The concept of secondary reinforcement has been extremely useful to psychological theorists and experiments centered around it have brought fruitful and important additions to our knowledge of behavior. Repeatedly, it has been demonstrated in animal experiments that stimuli which are paired with food or water will gain the power to "reinforce" behavior—either to sustain old responses or fixate new ones. However, while reinforcement theorists have generally made no distinction between the functional properties of food and water reinforcement and the reinforcement provided by the termination of noxious stimuli, almost all attempts to use the latter type of event as the basis for establishing secondary reinforcement have produced negative results. Yet, for example, when electric shock is used to motivate and reinforce behavior directly, learning is powerful and prompt.

If secondary reinforcement cannot be established with the termination of a noxious drive, there would seem to be little point to using drive reduction as the fundamental reinforcement mechanism in theoretical systems which at the same time lean heavily on secondary reinforcement—as, for example, Hull (1943), Mowrer (1956, 1959), and Miller (1951) have done. Indeed, the failure to demonstrate this phenomenon could be construed as presumptive evidence in favor of the opposite view: namely, the argument that secondary reinforcement is a phenomenon involving motivational increments, particularly those related to the stimulating properties of the anticipatory goal response. Spence (1956) and Seward (1952) have argued for this position.

In brief, then, while one would hesitate to mention such a will-o-the-wisp as a "crucial experiment," we cannot take lightly Mowrer's (1959) suggestion that the fate of drive reduction theory may rest on the successful demonstration of secondary reinforcement established in conjuction with the reduction of shock or some other noxious stimulation. The purpose of the present paper is to evaluate the evidence on the problem with an eye toward finding and/or proposing experimental tests of the hypothesis that secondary reinforcement can be so established.

EXPERIMENTAL EVIDENCE

For convenience the experimental evidence is categorized as follows, according to the method used in testing for reinforcement: (*a*) *response acquisition*, including bar pressing, T maze learning, head turning, and pushing a nose-key; (*b*) *response extinction;* (*c*) *delay of reinforcement;* (*d*) *response facilitation;* and (*e*) *preference testing* (other than T maze).

[1] Part of this material was originally presented in an unpublished doctoral dissertation, University of Illinois, 1958.

[2] Now at Wake Forest College, Winston-Salem, North Carolina.

The writer wishes to express his appreciation to O. H. Mowrer for critical discussions and comments on this material and to G. R. Grice, L. I. O'Kelly, and W. F. Crowder for their able suggestions.

Response Acquisition

Bar pressing. One of the first to claim positive results on this problem was Barlow (1952). In the two experimental groups of interest here, a 5-second light either came on (*a*) in the last 5 seconds of a 10-second shock, or (*b*) immediately after the shock. After a single such pairing each rat was tested 20 hours later, with total duration of bar pressing as the measure of reinforcement. For half the animals in each training group the light was continuously on and could be turned off by pressing the bar, while for the other half the light was off and could be turned on. The animals in Group *a* showed no significant difference in duration of bar pressing, though tending to turn the light on more than off. The animals in Group *b* showed a significant difference in this same direction. These results are very weak supporting evidence and give rise to two points of interest. First, secondary reinforcement was supposed to have been established with a single pairing. This result may be compared with those of Bersh (1951), who found that 80 light-food pairings did not produce a reinforcement effect significantly different from zero pairings. Second, significant results were found for the group in which the light came on after shock termination, but not for the group in which the light preceded it. The light in this instance could have come before, or have been paired with a discriminable drop in drive, thus accounting for the apparent backward conditioning.

But, it would appear from other sources that this particular sequence might bring about backward aversive conditioning, not positive conditioning. Razran (1956) concludes from his review of the literature that "with shock as the US, backward conditioning seems to be possible only when the CS is applied after the shock has ceased, and not when it is applied during the action of the shock." Mowrer and Aiken (1954) found that a signal following immediately upon the termination of shock later inhibited a food-reinforced response. It is consequently possible that in Barlow's experiment the light was aversive and when the subjects pressed the bar the onset of the light caused them to "freeze" on it, thus producing relatively long durations of pressing. Perhaps the duration measure is not the most satisfactory index of reinforcement.

Littman and Wade (1955) used a tail-shock apparatus to pair a light with shock termination. In a different apparatus, with light as the reinforcement for bar pressing, the rats did not press more than control subjects. Deutsch (1956; see also Littman & Wade, 1956) raised several questions about this experiment, the most pertinent of which concerns the use of a different apparatus for testing than that used for training. Direct evidence regarding the potency of a secondary reinforcer in transituational testing is not plentiful, and it is probably unwarranted to *assume* secondary reinforcement should be obtainable in a situation very different from that in which the subjects are trained, even with appetitive reinforcement.

Beck (1958) trained rats to escape from grid shock with a lighted T maze door and a tone serving as cues. After 180 training trials there were two 10-minute test periods during which the subjects were locked in the choice area of the maze with a newly-introduced bar. Shock was on continuously during testing and escape was not permitted. When the bar was pressed by the subjects in the experimental group the light and tone came on for 2.5 seconds. In the first of two replications of this experiment there was some indication that the light and tone were reinforcing bar pressing, but the effect was not strong and did not appear in the second replication. To further complicate the results, animals trained to escape the shock without any light or tone made as many or more bar presses with these as "reinforcers" as did the main experimental group.[3] A major difficulty in using grid shock is that the animals can get "primary" reinforcement by hitting upon any movement or posture which reduces the pain. Such responses can either compete with the test response or facilitate it, thereby increasing variability and possibly eliminating significant differences which might otherwise have been obtained.

T maze learning. The paradigm for the T maze experiments is fundamentally the same as Saltzman's (1949) experimental design. After rats have been reinforced in a distinctive goal box at the end of a straight runway, the reinforcing properties of the goal box are tested by putting it on one arm of a T maze, comparing turns to this box by experimental and control groups.

Smith and Buchanan (1954), used an amount-of-reinforcement design with this paradigm. Omitting the various controls, they trained one group of rats to run across an electrified grid to get food in a black goal box and across a sponge runway to get food in a white goal box. A second group had the color of the goal boxes reversed. The goal box associated with both food and shock reduction should take on greater secondary reinforcing capacity than the goal box associated with food alone. In a black-white discrimination situation, with the black goal box of a T maze positive, it was predicted that the animals previously shocked prior to entering the black goal box should make fewer errors than the group shocked prior to entering the white goal box. The results bore out the prediction quite well.

In three later experiments with the same basic design, Buchanan (1958) found that (*a*) rats would "increase their tendency to approach cues contiguous with escape from a fear-producing situation, as well as those contiguous with escape from shock"; (*b*) "the approach tendencies, acquired by hungry rats during training to cues associated with shock reduction and hunger reduction, were not appreciably affected by changes in the drive conditions of hunger and fear between training and testing"; and (*c*) "shock reduction and hunger reduction were approximately equal in their effects on the strength of acquired tendencies to approach associated cues, and that the drives of hunger and shock and/or their respective incentives combine in some fashion in the development of these approach tendencies" (p. 362).

In a similar kind of study, Nefzger (1957) trained rats to run across a grid into a distinctive end box. He hypothesized that as training progressed, the animals should show an increasing preference for this end box if it were on one arm of a T maze. He recorded no change of preference with repeated testing, however, and

[3] In this experiment secondary reinforcement was not shown using the same procedure and animals trained with water reinforcement and tested under 23 hours' water deprivation. The possible reasons for this, as well as pertinent data, are presented at length in the original report.

was unable to duplicate the Smith and Buchanan results when response elicitation by the goal box itself was controlled.

The problem of controlling for response elicitation during tests for secondary reinforcement is important for a number of experimental designs which utilize the same response in both training and testing.[4] By "elicitation" we refer to the capacity of a stimulus to evoke or otherwise exert discriminative control over a response. The "reinforcing" property of a stimulus refers to its capacity to be effective in fixating or otherwise to prolong responding in some manner. The question raised by an experiment such as Smith and Buchanan's is whether the black goal box in the T maze is eliciting an earlier-learned approach response (the goal box being visible from the choice point), or whether it is reinforcing some new turning response. If we are *testing* the hypothesis that reinforcement can be demonstrated in such a situation we must accept the eliciting interpretation over the reinforcing interpretation in case of any doubt. Buchanan, in his later experiments, did in fact abandon secondary reinforcement as the sole interpretation of his results and referred to the "eliciting and/or reinforcing" properties of the goal box. McGuigan (1956) has also discussed this general problem in relation to Hull's treatment of secondary reinforcement.

Head turning. Coppock (1950, 1951) obtained results indicative of the establishment of secondary reinforcement, pairing a blinking light with tail-shock termination. In his tests, whenever the rats had their heads turned 22° to a particular side they were continuously reinforced by the blinking light. The results were weakly positive, however, only for those animals which (*a*) had the light following shock termination, and (*b*) were reinforced with the head on the initially nonpreferred side. The effect thus appears to be very unstable, if real at all, and the use of the duration measure of head position is open to the same ambiguity as in Barlow's experiment, i.e., the blinking light might be producing fearful "freezing" of the head in the position.

Key-nosing. Crowder (1958) used a tail-shock apparatus with rats in several different experimental designs, all of which were characterized by very precise control of the shock, conducting tests for secondary reinforcement with the shock on, and using the pushing of a nose-key (similar to a Gerbrands pigeon key) in the front of the apparatus as an operant response. He found in one study that a light repeatedly paired with the termination of inescapable shock did not later have a significant reinforcing effect on the nosing response.

Response Extinction

The familiar model for this type of design is the comparison of rate of extinction with and without the presentation of a secondary reinforcer following responses during extinction. Bugelski's (1938) experiment is the prototype.

Crowder (1958), with his tail-shock and nose-key apparatus, gradually increased shock to a maximum intensity over a 25-second interval. The first nosing response after the shock reached its peak was immediately followed by a 0.5-second pre-

[4] Here we are concerned with the logical problem of interpreting experimental results in an ambiguous situation. In a later section we shall consider in detail the evidence related to the so-called "discriminative stimulus hypothesis" of secondary reinforcement.

sentation of light, then the shock was terminated. In both extinction and reconditioning the presentation of this light as a reinforcer significantly increased response rate. However, a modification of this procedure in which the shock during training came on instantaneously to full intensity produced completely negative results. Crowder's positive results seem to be the best indication of secondary reinforcement among all the studies reviewed, and his technique of gradually increasing shock intensity appears promising, though possibly having little effect other than adapting the animals to the pain.

Murphy, Miller, and Brown (1958) studied the extinction of an avoidance response. During training a light followed barrier-jumping responses to the shock and the CS. In extinction, with only the CS presented, it was found that the presentation of the light after each response prolonged extinction very markedly and the authors interpreted this to mean that secondary reinforcement had been demonstrated with pain and fear reduction as primary reinforcers. On the other hand, we are again faced with the awkward problem of backward conditioning (the light followed response and reinforcement during training) and the criticisms of Barlow's experiment hold here, also. It seems just as reasonable to argue that the light had become *aversive* during training and retarded extinction by keeping the general level of fear high rather than serving as positive reinforcement. A paper by Seeman and Greenberg (1955) is directly relevant to this experiment, as well as a brief report by Bender (1955).

Beck[5] trained five rats to escape from shock by running through the lighted member of a pair of adjoining plexiglass panels. Each animal was then put into the apparatus for 15 minutes with both panels locked and dark (no shock). Pushing against one of the panels always produced for 0.5 second the light which had been the positive stimulus, but escape was not permitted. Pushing against the other panel did not produce any illumination change. As predicted, (*a*) the subjects pushed the panel on the reinforced side significantly more than on the nonreinforced side, and (*b*) the percentage of total responses to the reinforced side was significantly greater than that of a control group trained without a cue stimulus. It is still possible, however, that when the light came on it was eliciting further responses rather than reinforcing a certain position habit.

Delay of Reinforcement

In a fourth experiment, Crowder (1958) used a light to bridge a delay between the nosing response and shock termination. Shock came on immediately at full intensity, then when the rat pushed the nose-key there was a 2-second onset of light, following which the shock was terminated. Eighty-five trials with this procedure did not result in significantly shorter latencies in occurrence of the response after shock onset than did a 2-second delay of reinforcement without the light. Both conditions were vastly inferior to immediate reinforcement.

Response Facilitation

Lee (1951) taught three groups of rats to bar press for food, then associated a tone with shock in a tail-shock apparatus. In one group the tone was associated with the onset of shock, with the termination of shock

[5] Unpublished research, University of Illinois, 1957.

in a second group, and with no shock at all in a third. These groups were now tested by pairing the tone with the previously-learned bar pressing response. According to his hypothesis, the group having tone associated with shock termination should press the most; the group without tone-shock experience an intermediate amount; the group with tone associated with shock onset should press the least. As it happened, pairing the tone with shock termination inhibited responding as much as pairing it with shock onset.

Mowrer and Aiken (1954) obtained results similar to Lee's. They paired a blinking light with shock onset and termination in various temporal sequences and found that after the light had been contiguous with shock termination (either just before or just after) its presentation inhibited bar pressing for food reinforcement. Mowrer and Aiken suggest that the light did not facilitate bar pressing because the animals were not afraid at the time of testing, i.e., the relevant motivating condition was not operative at the time of testing.

Preference Testing

With this paradigm, a distinctive environment is associated with escape from shock. This environment is then matched with some other stimulus context and the subjects' preferences are recorded.

After training their animals to escape from shock by running to a non-shock escape chamber, Goodson and Brownstein (1955) found that the escape chamber was preferred to either the shock compartment or a neutral compartment. Again, however, the test situation was so contrived that the results can be interpreted in terms of the elicitation of the previously learned escape response. During training the animals learned to run away from the shock box and into the escape box. Both of these aspects of running were reinforced by shock termination. In testing, the animals were put into an alleyway between two closed doors. These doors were simultaneously raised so that the animal was faced with the same situation encountered in training: an open door into the escape chamber, with all its eliciting cues, was present. The response scored, running into the escape compartment, was exactly the same response on which the animal had been trained.

Montgomery and Galton (1956) eliminated the testing ambiguity found in the Goodson and Brownstein experiment, but introduced another confusion. After placing a rat in a small plexiglass "trolley car," in one apparatus compartment shock was turned on and remained on while the animal was pulled into a second compartment, where it terminated. After a number of such trials the animal was put into the two-compartment situation for unrestricted running and time spent in the two compartments was recorded. Unfortunately, the fact that the subjects preferred the side where shock terminated can be interpreted to mean that they were *avoiding* the fear-arousing side. This is a phenomenon so well-established that there is no necessity for talking about secondary reinforcement.

A preference-testing experiment which would give clear-cut results would combine the *acquisition* procedure used in the Montgomery and Galton experiment and the *test* procedure of the Goodson and Brownstein experiment. Since in Montgomery's procedure the animals are *transported* from one compartment to another no particular running response is learned: hence, such a habit could not manifest itself during testing and be confounded with a demonstration of secondary reinforcement. Goodson's test procedure of pairing a neutral box with both the shock box and escape box in preference tests should show whether the animals are simply avoiding the shock compartment or have learned a preference for the escape side. Gleitman (1955) reports a study which is, in principle, the same as this "ideal" experiment. Rats were placed in a transparent cable car with a grid floor. Shock was turned on in one part of the experimental room, continued while the rats were transported via cable to another part of the room, then turned off. The subjects were divided into three groups for the testing of preferences and it was found that the termination place was preferred to the shock-onset point, there was no preference between a neutral place and termination point, and there was no preference between a neutral place and shock-onset locus. The ambiguity in these results is the failure to show a clear approach *or* avoidance tendency in the groups having the neutral place as a choice. Two procedural aspects of the experiment which might weaken any interpretation placed on it are the facts that (*a*) because the animals were run in an "open" room there may have been present what Mowrer (1959) has called "ambiguous cues," stimuli associated with both the onset and termination of shock; and (*b*) there were 11 experimenters, students in an experimental laboratory. As we shall see later, however, even a perfectly controlled experiment with this design might fail to give positive results, for it may not be possible to establish secondary reinforcement without the animals making a discriminative response during training.

Summary

The experimental literature shows a variety of tests of the hypothesis that the termination of aversive stimulation can be used as the primary reinforcer in establishing a secondary reinforcer, but there are few positive results. In those instances where there has been a clear experimental effect the interpretation is generally confounded such that the concept of secondary reinforcement need not be invoked. Only one experiment (Crowder) seems to be unambiguously positive, with a highly significant effect, but in view of his own negative results in other experiments, as well as the rest of the literature, this does not provide an undue amount of faith that the phenomenon exists. We must now ask whether this predominantly negative evidence is sufficiently strong to refute the theoretical positions which predict that secondary reinforcement should be established under such conditions or whether the explanations for the experimental failures are to be found in the experiments themselves. Toward this end a consideration of two variables related to secondary reinforcement becomes appropriate: namely, discrimination training and motivation.

Secondary Reinforcement and Discrimination Training

Discriminative Stimulus Hypothesis of Secondary Reinforcement

One of the difficulties involved in trying to interpret the experimental literature in the previous section was the lack of differentiation between the eliciting and reinforcing functions of stimuli. There, we considered only the logical problem that in many situations the repeated occurrence of a response could be attributed to its elicitation by some previously-learned cue and that the concept of secondary reinforcement was therefore superfluous. The black goal box visible from the choice point of a T maze is a case in point. The question to be considered now is somewhat different, to wit: what is the empirical relationship between the eliciting and reinforcing functions of stimuli? Specifically, can a stimulus serve as a secondary reinforcer without first having cue function?

Keller and Schoenfeld (1950, p. 236) have stated this discriminative stimulus hypothesis in firm terms: "*In order to act as an* S^r *for any response a stimulus must have status as an* S^D *for some response.*" This cue function is established through a process of differential reinforcement, reinforcing a response in the presence of a stimulus and not in its absence. The importance of such a procedure in discrimination training has been shown by Ferster (1951), for example, who found that a stimulus continuously present during both reinforcement and nonreinforcement did not have the properties of a discriminative stimulus, i.e., had no control over behavior. In order to clarify the nature of the establishment of a discriminative stimulus, its use as a secondary reinforcer, and the nature of the problem of the interaction of these two properties, a brief review of the typical Skinner-box training procedure is in order before examining experimental results.

The procedure is generally as follows: (*a*) An animal is trained to eat from a food magazine. (*b*) It is trained to eat in the presence of a certain stimulus, such as a light, or immediately following some stimulus, such as the click of the food delivery mechanism. Such stimuli are referred to as *discriminative stimuli* (symbolized by S^D) and are the same as the positive stimuli in any discrimination training situation. When the discriminative stimulus is not present, nor has just occurred (depending on the kind of training situation), the animal is not reinforced for going to the magazine. (*c*) A bar is introduced for the animal to press. As soon as it is pressed, S^D follows immediately and the animal can go to the magazine and eat. The animal is then extinguished on bar pressing, with or without the S^D following the response, and the occurrence of S^D is found to increase resistance to extinction. As an alternative procedure the animal may originally learn to press the bar with only the S^D as reinforcement. Under either of these conditions the discriminative stimulus is referred to as a secondary reinforcer.

Most of the experimental tests of the discriminative stimulus hypothesis have been positive. Schoenfeld, Antonities, and Bersh (1950) found that the mere temporal contiguity of a stimulus with some reinforcer was not sufficient to establish this as a secondary reinforcer (compare with Ferster above, also). In their study, a light was associated with the consumption of food pellets, but not with obtaining them. After 100 such associations, the light did not increase the rate of bar pressing above operant level.

Dinsmoor (1950) studied the discriminative stimulus—secondary reinforcement relationship with an extinction procedure. After training rats on bar pressing, half were extinguished with the presentation of S^D as reinforcement following responses. For the other half, the bar was removed and the experimenter presented the S^D without food the same number of times that it had occurred in the first group. When the bar was again made available to the second group it was found to have extinguished on bar pressing as much as the first group, i.e., response rate was reduced the same amount. Extinction was carried further, with "cue" and "reinforcing" functions of the S^D reversed for the two groups and no differential rate of extinction was found now either, a clear demonstration of the intimacy of these two properties of stimuli. Coate (1956) replicated part of Dinsmoor's experiment with the same results.

Notterman (1951) studied secondary reinforcement as a function of amount of discrimination training. Interspersing a varying number of nonreinforced trials in which S^D was absent among a constant number of reinforced trials in which S^D was present, this investigator found that the more strongly the discrimination was thus developed the greater secondary reinforcing power the S^D had. McGuigan and Crockett (1957, 1958), Webb and Nolan (1953), and Wike and McNamara (1956) report similar results.

These experiments indicate, then, that (*a*) with better discrimination training secondary reinforcement is stronger, and (*b*) when either the cue or reinforcing function of a stimulus is extinguished, the alternate function also declines. The rationale for maintaining the distinction then is the way in which the stimulus is used, its temporal relationship to a particular bit of behavior. Keller and Schoenfeld have clearly made this point, also. Contrary to these positive results, however, there are three reports of experiments which apparently contradict the hypothesis.

Rozeboom (1957) was unable to replicate the Dinsmoor-Coate results, and even got little depression of bar pressing after pairing the S^D with shock during the phase when the response to the S^D was being extinguished. He does not report any data from the extinction period, unfortunately. In both Rozeboom's experiment and Wyckoff's (reported below) a somewhat unusual procedure was used. Rather than food reinforcement, water reinforcement was used, delivered by a dipper which was normally up and lowered into a reservoir for water at the appropriate time. Rozeboom's latent extinction procedure involved having the dipper mechanism operate without water during the cue extinction period. The subjects could still make the licking response to the dry dipper during the latent extinction period.

Wyckoff, Sidowski, and Chambliss (1958) trained their rats to approach and lick the dry dipper when a buzzer sounded, whereupon the dipper delivered water. After this training, a bar was inserted into the side of the box opposite the dipper and animals for whom the buzzer was contiguous with bar presses failed to make more responses than animals for whom the buzzer sounded automatically following 10 seconds of no bar pressing, the dipper no longer operating in either case. While the authors themselves felt "no inclination to reject the concept of secondary reinforcement," they did believe that some crucial condition in the establishment of secondary reinforcement remains unspecified. In this experiment, the buzzer was not *directly* associated with water presentation and a consummatory response (licking water from a dipper), but instead was paired with an operant response (licking dry dipper), which was similar to the consummatory response. Perhaps the crucial difference sought for by Wyckoff is related to this fact—a suggestion made promising by Rozeboom's negative results with the dry dipper technique.

On the basis of the Wyckoff study, Myers (1958) suggests that many bar pressing experiments thought to have shown secondary reinforcement may have really shown only heightened activity following the presentation of S^D. When this was controlled, "secondary reinforcement" no longer appeared. Direct experimental evidence in support of this contention (which Myers did not present) can be found. Both Walker (1942) and Estes (1948) found that when an S^D established under Condition *X* was periodically presented under Condition *Y* it increased the rate of a response with which it had never before been directly associated. Gilbert and Sturdivant (1958) report similar results. It would not be surprising then to find that an S^D would facilitate a response in the same situation where it had been associated with that response. Zimmerman (1957, 1959), however, contrary to Wyckoff, has obtained literally thousands of responses from his animals with nothing but secondary reinforcement. In the same kind of control group as used by Wyckoff responding was very low, no more than operant level. It would seem valuable to repeat the Wyckoff experiment to ascertain just what variables are operating. Wyckoff himself seems not to be dismayed by it all, for he has since attempted to develop a quantitative theory of secondary reinforcement (1959) using "cue strength" as the main variable influencing the secondary reinforcing capacity of a stimulus. In any event, these results would not seem to influence the interpretation of experiments using such apparatuses as straight runway or T maze, where trials are spaced.

The third contradictory report is Ratner's (1956). After training his rats to approach a hopper at the sound of a click he introduced a bar into the box. Animals for whom the bar pressing was followed by the click made more presses than a no-click group, but did not go to the hopper more often. Ratner suggests that although the click was reinforcing it was not an S^D for goal-approaching because the animals went to the goal box only a small proportion of the time that the click was presented (about 20% on the first day). Often, however, either rats or pigeons will not go to the goal after every response on a schedule of 100% primary reinforcement. In fact, one of the things we expect a reinforcer to do is to "strengthen a habit" so that the habit will maintain itself without external reinforcement. In addition, in Ratner's situation other cues, such as the sight or sound of the food being delivered, may also have been important as S^D in control of goal-approaching and these may have been absent during testing. This briefly reported experiment is suggestive, but inconclusive.

In view of the total evidence, it seems that something about the nature of discrimination training is important for the establishment of secondary reinforcement. While we cannot here go into the problem of the possible underlying mechanisms we are inclined to take the evidence at its face value. Since it may be possible to obtain secondary reinforcement without prior discrimination training and an S^D does not seem to be guaranteed as a reinforcer we can not accept cue function as prima facie evidence for secondary reinforcement. It seems clear, however that such training generally does make secondary reinforcement stronger and to this extent the S^D hypothesis may be considered as "correct," providing us with some empirical basis for analyzing the failure of some of the shock-termination experiments.

Discrimination Training and the Shock-Termination Problem

By and large, the shock-termination experiments have been based on a Hullian-type assumption that the sufficient condition for establishing secondary reinforcement is pairing a neutral stimulus with some "primary" or other "secondary" reinforcer (see McGuigan, 1956). There is no specific statement of prior discrimination training in this assumption (although Hull has written of the eliciting properties of the secondary reinforcer), and in none of the experiments thus far reported, save those of Smith and Buchanan and of Beck, has there been any attempt to use discrimination training. Therefore, it seems that the majority of negative and questionable results cannot necessarily be considered as evidence that the phenomenon does not exist, or as clearly opposing drive-reduction theory. Rather, they can be considered as evidence against only a particular statement as to how secondary reinforcement can be established in conjunction with drive-reduction reinforcement, i.e., that the simple pairing of a neutral stimulus and drive reduction is sufficient. If this statement is incomplete, and the evidence indicates that it is, then the experiments have hardly touched upon the main problem we are considering—whether secondary reinforcement can be established *at all* using shock termination as primary reinforcement. Granting that the procedures used with food or water may not necessarily be correct for use with the termination of noxious drives, they still provide the best direction for research.

Secondary Reinforcement and Motivation

Recalling Mowrer and Aiken's suggestion that perhaps they failed to obtain secondary reinforcement because their animals were not appropriately motivated during testing, we look back over the other experiments and see that only in Crowder's and Beck's experiments were the animals shocked during testing. On the other hand, in experiments with hunger and thirst the subjects are tested for secondary reinforcement under the same deprivation conditions with which they were trained. The exception to this occurs, of course, in the few experiments which have studied secondary reinforcement as a function of motivation.

Brown (1956) found in tests for secondary reinforcement that there was no interaction between hunger level and amount of responding in reinforcement and nonreinforcement groups. The secondary reinforcement group responded equally more than the control group at both "high" and "low" drive levels. She suggests that satiated animals getting the secondary reinforcer might very likely have not given any indication of effectiveness of the reinforcing stimulus, but her low-drive group was not satiated.

Miles (1956) obtained similar results. After training his rats on bar pressing under 23 hours' food deprivation he gave them extinction testing under 0, 2.5, 5, 10, 20, and 40 hours of deprivation. At each drive level the experimental group was superior to a comparably trained and deprived control group which did not get the secondary reinforcer. Like Brown, he found no regular trend for the difference between experimental and control groups to increase as a function of deprivation time, although the three shortest deprivation groups showed less difference than the three highest. The experimental-control differences were not significant at the shorter deprivation intervals, but the overall functions were.

Oakes (1956), on the other hand, did find an interaction. Varying S^D presentation and food deprivation time in a factorial design, he found that both of these variables influence straight runway performance. His high-drive group with cue reinforcement ran faster than either the low-drive group with cue or the high-drive group without cue reinforcement.

Wike and Casey (1954) claim to have demonstrated the secondary reinforcing property of food for satiated animals, finding that the satiated animals which got food (which they did not eat) in the goal box ran a straight runway faster than rats which did not get pellets in the goal box. Unfortunately, as these writers themselves report, there was no control for the effect of simply manipulating something in the goal box, for example, nonedible objects.

Schlosberg and Pratt (1956) got very marked results indicating the difficulty of demonstrating secondary reinforcement with satiated subjects. Rats under 23 hours' deprivation showed a consistent preference for the side of a T maze where they could see and smell food, but not eat it. When run while satiated the preference reduced to chance, only to return immediately to its former high level when the rats were again deprived. Rats initially run in the maze while satiated showed only chance preferences, and when switched to deprivation took as long to display the preference as did the first group, run deprived from the start. The authors conclude that hunger was necessary for both *learning* and *maintaining* the preference.

In a recent study Grice and Dyal[6] gave three groups of rats 110 click-food pairings per animal while the subjects were under 23 hours' food deprivation. After this training a bar was introduced into the apparatus and the animals were tested for 30 minutes. There was a very significant reinforcement effect between a 23-hour deprived-click-reinforcement group and a 23-hour no-click group, the means being 56.12 and 17.12. However, a satiated-click group gave almost twice as many responses as the hungry-no-click group, a mean of 30.88. This suggests that while secondary reinforcement may be obtainable with satiated subjects, it is certainly more powerful with deprived.

Seward and Levy (1953) obtained results difficult to interpret. Rats given food reinforcement on one side of a T maze continued to show a preference for this side when run satiated. On the other hand, with repeated training and testing they showed no increasing preference for the food side as would be expected if secondary reinforcement were operating.

These various experiments on the relationship of drive level and secondary reinforcement, while few in number and sometimes contradictory, strongly suggest that any such effect obtained with satiated animals would be weak. Only Grice and

[6] Unpublished research, University of Illinois, 1957.

Dyal obtained a difference even close to significance with satiated animals, excepting the doubtful results of Seaward and Levy and of Wike and Casey. If we now assume that motivation is an important variable and look at the problem of the shock-termination experiments again we first have to determine the condition equivalent to "satiation."

Satiation in Shock Experiments

As the term is generally used, "satiation" refers to the absence of some motivating condition, of the "complete gratification" of a need. In practice, this means that we have induced an animal to eat or drink as much as it will so that we are reasonably assured that it does not "need" food or water. In the shock situation satiation should then refer to the absence of shock. There is another component of shock situations, however, namely, fear. Thus, even though shock were not present while testing in the same apparatus used in training there might be considerable motivation and we would not think of the subjects as satiated. Whether this fear component would be a powerful enough motivator to demonstrate secondary reinforcement clearly, assuming its demonstrability, would presumably be dependent upon such parameters as strength and number of shocks during training. Since Miller (1951) has shown that animals will learn a variety of responses with escape-from-fear reinforcement, we might expect that this same motivation would be potent enough to demonstrate secondary reinforcement. Schoenfeld (1950) believes that it can be so demonstrated, having hypothesized that in avoidance learning the animals continue to make the avoidant response because the proprioceptive stimuli associated therewith have taken on secondary reinforcing properties.

We might get satiation on the other hand, if we tested the subjects in a completely different apparatus, as, for example, in the Littman and Wade experiment. Mason,[7] however, has made the suggestion that in this case the stimulus which was supposed to be reinforcing might have the opposite effect and arouse fear, inhibiting responses on account of its prior association with the shock situation. This is an especially interesting argument, for its implication to the drive reductionist is that a signal associated with shock termination will arouse drive in a nonshock situation and reduce it when the organism is pained or fearful. One would not then expect to get positive results in a situation like that of Littman and Wade, but would expect results like those of Lee and of Mowrer and Aiken.

Effects of Very Strong Shock During Testing

Thus far we have been considering the effects of very low motivation or complete absence of motivation during tests for secondary reinforcement. What about the converse, can there be too much motivation during testing? None of the hunger or thirst experiments concerned with secondary reinforcement and drive have apparently had the subjects too highly motivated, but shock can easily produce an excitation level far beyond that of appetitive drives and has been shown to have ill-effects on performance (for example, as far back as Yerkes & Dodson, 1908). How does this deleterious effect of very strong motivation apply in the shock

[7] D. J. Mason, Personal communication, 1956.

situation we have been considering.

Competing responses. As mentioned previously, shock may arouse responses in competition with or facilitating the response being measured. The combination of these opposite effects in a single experiment would tend to wash out experimental differences.

Psychophysics of drive-reduction reinforcement. A second problem may be even more serious since it strikes at the nature of the mechanism of secondary reinforcement. Let us assume momentarily that the basis of secondary reinforcement is some form of anticipatory drive reduction. Campbell and Kraeling (1953) have shown that the effectiveness of shock reduction as a reinforcer is a function of the *proportion* of the total shock reduced, not just the absolute amount of reduction. This approximates a Weber fraction, which Campbell (1955) has shown even more clearly with sound reduction. If, then, an animal is being shocked during a test for secondary reinforcement, the amount of anticipatory drive reduction induced by a secondary reinforcer could well be less than the differential threshold for reinforcement. What might potentially be a "good" secondary reinforcer could reduce such a small proportion of the total drive in this situation as to be ineffective. We could therefore tell if the reinforcement "threshold" were reached only if reinforcement were demonstrated. If it were not demonstrated one could argue that the threshold had not been reached and the hypothesis was not disproved at all. The way to break out of this circularity would seem to be a careful study of a variety of shock and/or fear levels. The a priori selection of a shock level for testing would not seem to be adequate, even though the test shock were the same as in training, because the termination of a given shock intensity might be an effective reinforcer for discrimination training but the shock inappropriate for continual use during tests for secondary reinforcement.

Motivation During Training

The role of motivation during discrimination training is not considered in detail because one can observe and measure discrimination performance with enough accuracy to tell when a discriminative response has been well-learned. There is also good evidence that shock intensity is relatively unimportant if the discrimination to be learned is simple (for example, Hammes, 1956).

Summarizing these various lines of evidence, then, it may be suggested that some amount of aversive motivation will probably be necessary to demonstrate secondary reinforcement derived through association with pain reduction. The precise level is a matter to be determined empirically, but intensities either too high or too low may negate otherwise positive results.

Some Suggested Experimental Approaches

Only Crowder, to the writer's knowledge, has really attempted the "obvious" experiment, using the design that Bugelski used over 20 years ago—extinguishing a response with and without the use of an hypothesized secondary reinforcer. Crowder's results suggest that this might be a fruitful approach, especially since he apparently did get positive results even without discrimination training. A powerful procedural advance would be an extensive use of partial reinforcement, particularly the method that Zimmerman (1957, 1959) has reported for use with food and water reinforcement. With this technique, the S^D follows bar pressing only part of the time, and reinforcement follows the S^D only a fraction of the time that it occurs. Using grid shock in conjunction with a discrimination-training procedure and bar press training, one might produce satisfactory secondary reinforcement during extinction with the cue stimulus as the reinforcer.

Fear motivation might be even more effective than direct shock, for many of the problems encountered with grid shock could be avoided. Suppose that we put a rat into a shock compartment with a hinged door in one of the walls. We block off this door and shock the animal severely, à la Miller. Now we turn off the shock and make the door available, gradually training the animal in repeated trials to push open the door and run out of the shock compartment to a safe compartment. We run 10 such trials a day, shocking the animal before training each day to insure that the level of fear is high. We now introduce an S^D, such as a buzzer. The rat is put in the box and the panel is locked, not to be opened until the buzzer sounds. Soon the animal learns not to push the door until the signal is presented. Then, à la Zimmerman, we put this panel pushing on a partial schedule, such that when the buzzer sounds the animal does not always get to escape when he pushes on cue. Rather, the buzzer ceases when the animal responds, then comes on again after another minute or so. We slowly build up this ratio so that the buzzer sounds several times before the animal finally is reinforced. Now comes the test period. Everything is the same on this day except that there is a bar in the box, which when pressed sounds the buzzer for a period of time, but the animal is not allowed to escape. We can thus test the reinforcing capacity of the buzzer in the same manner that we test for secondary reinforcement with appetitive motivation. In the same analogous manner we could have introduced the bar during training itself, then tested for secondary reinforcement by presenting the buzzer (but no escape) during extinction. The critical aspects of either of these designs, in view of the foregoing discussions, are that (*a*) the buzzer is established as a discriminative stimulus, contiguous with escape from a fear-arousing situation, and (*b*) the motivational conditions during testing are the same as those during training.

There are a variety of aversive stimulus situations which could be used as alternatives to shock and fear. In a report concerning the use of cold stimulation and heat reinforcement, for example, Carlton and Marks (1957) report that it was very difficult to establish a stable bar pressing rate unless a cue stimulus preceded the onset of the heat. They interpret this to mean that the cue is serving as a secondary reinforcer. While this may not be a stringent test of reinforcement, the technique does seem readily amenable to more direct tests. One might use the cessation of strong light or sound as a reinforcer in the same way. Air deprivation and reinforcement would provide an interesting test, but an extinction procedure with the subjects deprived would rapidly become confounded.

Conclusions

This review of the experimental literature leads us to conclude that there is almost no evidence to show that secondary reinforcement can be established by the association of a neutral stimulus with noxious-drive reduction. An analysis of the experiments suggests that they have not been completely adequate for a variety of reasons, depending upon the particular experimental designs used. Generally speaking, there have been three major problems. First, there has often been a lack of certainty whether stimuli were eliciting previously-learned responses, or reinforcing them during tests for reinforcement. Second, in almost none of the experiments has the secondary-reinforcer-to-be first been established as a cue, although the literature strongly suggests that this procedure is advisable. Third, there has been relatively little consideration of the role of motivation during tests for secondary reinforcement.

It would seem that a first step in attacking this problem is to design experiments which provide maximal opportunity for the phenomenon to be demonstrated. It should be determined whether secondary reinforcement can be established at all, using methodologically sound and unambiguous procedures, before going on to test different hypotheses about the establishment of secondary reinforcement. Until this is done it seems meaningless to use the negative results thus far obtained as evidence against a concept as general as drive reduction. In the event that the phenomenon can never be demonstrated we may have a finding detrimental to drive-reduction theory, but certainly not to reinforcement theory since there are a number of alternative explanations for the operation of reinforcement. Reinforcement theory may be forced into the acceptance of some kind of hedonic axiom, however, and agree with P. T. Young that getting rid of something bad is not the same as getting something good.

REFERENCES

Barlow, J. A. Secondary motivation through classical conditioning: One trial nonmotor learning in the white rat. *Amer. Psychologist*, 1952, **7**, 273. (Abstract)

Beck, R. C. Secondary reinforcement and shock-motivated discrimination. Unpublished doctoral thesis, University of Illinois, 1958.

Bender, H. Relative effectiveness of start-box and goal-box cues in the maintenance of avoidance responses. *Amer. Psychologist*, 1955, **10**, 409–410. (Abstract)

Bersh, P. J. The influence of two variables upon the establishment of a secondary reinforcer for operant responses. *J. exp. Psychol.*, 1951, **41**, 62–73.

Brown, Janet L. The effect of drive on learning with secondary reinforcement. *J. comp. physiol. Psychol.*, 1956, **49**, 254–260.

Buchanan, G. The effects of various punishment-escape events upon subsequent choice behavior of rats. *J. comp. physiol. Psychol.*, 1958, **51**, 355–362.

Bugelski, R. Extinction with and without sub-goal reinforcement. *J. comp. Psychol.*, 1938, **26**, 121–133.

Campbell, B. A. The fractional reduction in noxious stimulation required to produce "just noticeable" learning. *J. comp. physiol. Psychol.*, 1955, **48**, 141–148.

Campbell, B. A., & Kraeling, Doris. Response strength as a function of drive level and amount of drive reduction. *J. exp. Psychol.*, 1953, **45**, 97–101.

Carlton, P. L., & Marks, R. A. Heat as a reinforcement for operant behavior. USA Med. Res. Lab., 1957, Project No. 6–95-20-001, Subtask, Climatic effects on psychophysiological abilities.

Coate, W. B. Weakening of conditioned bar-pressing by prior extinction of its discriminated operant. *J. comp. physiol. Psychol.*, 1956, **49**, 135–138.

Coppock, H. An investigation of the secondary reinforcing effect of a visual stimulus as a function of its temporal relation to shock termination. Unpublished doctoral dissertation, University of Indiana, 1950.

COPPOCK, H. W. Secondary reinforcing effect of a stimulus repeatedly presented after after electric shock. *Amer. Psychologist*, 1951, **6**, 277. (Abstract)

CROWDER, W. F. Secondary reinforcement and shock termination. Unpublished doctoral thesis, University of Illinois, 1958.

DEUTSCH, J. A. A note on the paper by Richard A. Littman and Edward A. Wade entitled "A negative test of the drive-reduction hypothesis." *Quart. J. exp. Psychol.*, 1956, **8**, 185.

DINSMOOR, J. A. A quantitative comparison of the discriminative and reinforcing functions of a stimulus. *J. exp. Psychol.*, 1950, **40**, 458–472.

ESTES, W. K. Discriminative conditioning: II. Effects of a Pavlovian conditioned stimulus upon a subsequently established operant response. *J. exp. Psychol.*, 1948, **38**, 173–177.

FERSTER, C. B. The effect on extinction responding of stimuli continuously present during conditioning. *J. exp. Psychol.*, 1951, **42**, 443–449.

GILBERT, T. F., & STUDIVANT, E. R. The effect of a food-associated stimulus on operant-level locomotor behavior. *J. comp. physiol. Psychol.*, 1958, **51**, 255–257.

GLEITMAN, H. Place learning without prior performance. *J. comp. physiol. Psychol.*, 1955, **48**, 77–79.

GOODSON, F., & BROWNSTEIN, A. Secondary reinforcing and motivating properties of stimuli contiguous with shock onset and termination. *J. comp. physiol. Psychol.*, 1955, **48**, 381–386.

HAMMES, J. A. Visual discrimination learning as a function of shock-fear and task difficulty. *J. comp. physiol. Psychol.*, 1956, **49**, 481–484.

HULL, C. L. *Principles of behavior.* New York: Appleton-Century-Crofts, 1943.

KELLER, F., & SCHOENFELD, W. *Principles of psychology.* New York: Appleton-Century-Crofts, 1950.

LEE, W. A. Approach and avoidance to a cue paired with the beginning and end of pain. Cited by D. C. McClelland, J. W. Atkinson, R. Clark, & E. L. Lowell in *The Achievement Motive.* New York: Appleton-Century-Crofts, 1953. P. 74.

LITTMAN, R., & WADE, E. A negative test of the drive-reduction hypothesis. *Quart. J. exp. Psychol.*, 1955, **7**, 56–66.

LITTMANN, R., & WADE, E. Reply to the note by J. A. Deutsch. *Quart. J. exp. Psychol.*, 1956, **8**, 186.

MCGUIGAN, F. J. The logical status of Hull's principle of secondary reinforcement. *Psychol. Rev.*, 1956, **63**, 303–308.

MCGUIGAN, F. J., & CROCKETT, F. A test of the discriminative stimulus: Secondary reinforcement hypothesis. *Amer. Psychologist*, 1957, **12**, 469. (Abstract)

MCGUIGAN, F. J., & CROCKETT, F. Evidence that the secondary reinforcing stimulus must be discriminated. *J. exp. Psychol.*, 1958, **55**, 184–187.

MILES, R. C. The relative effectiveness of secondary reinforcers throughout deprivation and habit strength parameters. *J. comp. physiol. Psychol.*, 1956, **49**, 126–130.

MILLER, N. E. Learnable drives and rewards. In S. S. Stevens (Ed.), *Handbook of experimental psychology.* New York: Wiley, 1951. Pp. 435–472.

MONTGOMERY, K. C., & GALTON, B. B. A test of the drive-reduction explanation of learned fear. Cited by R. L. Solomon, & E. S. Brush in Experimentally derived conceptions of anxiety and aversion. In M. R. Jones (Ed.), *Nebraska symposium on motivation.* Lincoln: Univer. Nebraska Press, 1956. Pp. 212–305.

MOWRER, O. H. Two-factor learning theory reconsidered, with special reference to secondary reinforcement and the concept of habit. *Psychol. Rev.*, 1956, **63**, 114–128.

MOWRER, O. H. *Learning theory and behavior.* New York: Wiley, 1959.

MOWRER, O. H., & AIKEN, E. G. Contiguity vs. drive-reduction in conditioned fear: Temporal variations in conditioned and unconditioned stimulus. *Amer. J. Psychol.*, 1954, **67**, 26–38.

MURPHY, J. V., MILLER, R. E., & BROWN, E. Secondary reinforcement and avoidance conditioning. *J. gen. Psychol.*, 1958, **59**, 201–209.

MYERS, J. L. Secondary reinforcement: A review of recent experimentation. *Psychol. Bull.*, 1958, **55**, 284–301.

NEFZGER, M. D. The properties of stimuli associated with shock reduction. *J. exp. Psychol.*, 1957, **53**, 184–188.

NOTTERMAN, J. M. A study of some relations among aperiodic reinforcement, discrimination training, and secondary reinforcement. *J. exp. Psychol.*, 1951, **41**, 161–169.

OAKES, W. F. The relationship between the effectiveness of a secondary reinforcer and certain stimulus conditions. Unpublished doctoral dissertation, University of Minnesota, 1956. (*Dissertation Abstr.*, **16**, 7372)

RATNER, S. G. Reinforcing and discriminative properties of the click in the Skinner box. *Psychol. Rep.*, 1956, **2**, 332.

Razran, G. Backward conditioning. *Psychol. Bull.*, 1956, **53**, 55–69.

Rozeboom, W. W. Secondary extinction of lever-pressing behavior in the albino rat. *J. exp. Psychol.*, 1957, **54**, 280–287.

Saltzman, I. J. Maze learning in the absence of primary reinforcement: A study of secondary reinforcement. *J. comp. physiol. Psychol.*, 1949, **42**, 161–173.

Schlosberg, H., & Pratt, C. H. The secondary reward value of food for hungry and satiated rats. *J. comp. physiol. Psychol.*, 1956, **49**, 149–152.

Schoenfeld, W. N. An experimental approach to anxiety, escape, and avoidance behavior. In P. H. Hoch, & J. Zubin (Eds.), *Anxiety.* New York: Grune & Stratton, 1950. Pp. 70–99.

Schoenfeld, W. N., Antonitis, J. J., & Bersh, P. J. A preliminary study of training conditions necessary for secondary reinforcement. *J. exp. Psychol.*, 1950, **40**. 40–45.

Seeman, W., & Greenberg, I. Secondary reinforcement and acquired drives: A note. *J. genet. Psychol.*, 1955, **86**, 367–373.

Seward, J. P. Introduction to a theory of motivation in learning. *Psychol. Rev.*, 1952, **59**, 405–413.

Seward, J. P., & Levy, N. Choice-point behavior as a function of secondary reinforcement with relevant drives satiated. *J. comp. physiol. Psychol.*, 1953, **46**, 334–338.

Smith, M. P., & Buchanan, G. Acquisition of secondary reward by cues associated with shock reduction. *J. exp. Psychol.*, 1954, **48**, 123–126.

Spence, K. W. *Behavior theory and conditioning.* New Haven: Yale Univer. Press, 1956.

Walker, K. C. The effect of a discriminative stimulus transferred to a previous unassociated response. *J. exp. Psychol.*, 1942, **31**, 312–321.

Webb, W. B., & Nolan, C. Y. Cues for discrimination as secondary reinforcing agents: A confirmation. *J. comp. physiol. Psychol.*, 1953, **46**, 180–181.

Wike, E. L., & Casey, A. The secondary reward value of food for satiated animals. *J. comp. physiol. Psychol.*, 1954, **47**, 441–443.

Wike, E. L., & McNamara, J. J.. Some training conditions affecting secondary reinforcement. *Amer. Psychologist*, 1956, **11**, 432. (Abstract)

Wyckoff, L. B. Toward a quantitative theory of secondary reinforcement. *Psychol. Rev.*, 1959, **66**, 68–78.

Wyckoff, L. B., Sidowski, J., & Chambliss, D. An experimental study of the relationship between secondary reinforcing and cue effects of a stimulus. *J. comp. physiol. Psychol.*, 1958, **51**, 103–109.

Yerkes, R. R., & Dodson, J. Relation of strength of stimlus to rapidity of habit-formation. *J. comp. Neurol. Psychol.*, 1908, **18**, 459–482.

Zimmerman, D. W. Durable secondary reinforcement: Method and theory. *Psychol. Rev.*, 1957, **64**, 373–383.

Zimmerman, D. W. Sustained performance in rats based on secondary reinforcement. *J. comp. physiol. Psychol.*, 1959, **52**, 353–358.

(Received August 3, 1959)

Psychological Bulletin
1961, Vol. 58, No. 1, 46–54

REPEATED MEASUREMENTS DESIGNS AND COUNTERBALANCING

JOHN GAITO[1]
Wilkes College

The repeated measurements analysis of variance designs have been popular in psychological research for a number of years. The advantage of these designs has been mainly that of economy, relative to number of subjects (*S*s), but increased precision may result; the experimental error is reduced when variance due to *S*s is removed. The simplest design of this nature is that of the treatments × subjects design in which *n S*s receive all of *k* treatments. More complex designs involve the latin square and modified latin squares.

Whenever repeated measurements designs have been used the procedure has usually been to counterbalance the order of appearance of the treatments so as to avoid any practice or learning effect which may be present. In the simple case this involves having all *S*s take the treatments in one order and then reversing the order on subsequent trials (intrasubject counterbalancing), or some *S*s take the treatments in one order and other *S*s receive different orders of presentation (intersubject counterbalancing). A combination of these two procedures probably is used most frequently. However, the practice effect is not partitioned in these analyses. In the more complex designs, the investigator is able to separate a source of variation due to practice or order. In some cases the effect due to the order by treatment interaction can also be partitioned.

The repeated measurements designs have been considered by numerous individuals (e.g., Alexander, 1947; Gaito, 1958a, 1958b; Garrett & Zubin, 1943; Grant, 1944, 1948; Gourlay, 1955; Hilgard, 1951; Kogan, 1948, 1953; Lindquist, 1947, 1953; Lubin, 1954, 1957, 1958; McNemar, 1951, 1955; Peters, 1944). Likewise, the criticisms directed at these designs have been numerous. The major indicated defect in the simple design is that the treatment effect will be confounded with any practice effect or, if counterbalancing is used, the main effects will be balanced but any practice effect will appear in the interaction effect, thus producing a negative *F* test bias (Type II error). The use of latin squares has been criticized because it has been maintained that interactions must be zero for valid use of this design. However, papers by Gourlay (1955) and Gaito (1958b) have indicated that this assumption is not always required. The latter individual employed the expected value of mean square [E(*MS*)] concept and showed that the important consideration as to the suitability of the latin square model depends on the number of random variates included in the experiment. Work by mathematical statisticians (Wilk & Kempthorne, 1957) has also indicated that interactions do not necessarily have to be zero.

The overall problem of repeated measurements designs is a complex one, and a satisfactory treatment has not been effected. However, the E(*MS*) concept (Anderson & Bancroft, 1952; Cornfield & Tukey,

[1] Now at Lake Forest College.

1956; Greenwood, 1956; Kempthorne, 1952; Wilk & Kempthorne, 1957) provides a suitable technique for a clear investigation of this problem. The purpose of this paper is to extend this approach to a number of repeated measurements designs. To investigate adequately this problem we shall treat six cases which are most frequently used: (*a*) all *S*s receive the Treatments (*T*) in the same Order (*O*); (*b*) the Order of Treatments is randomized for each *S*; (*c*) the Order is balanced (assumption that all interactions containing order are zero); (*d*) the Order is balanced and analyzed as a single latin square (no assumptions about interaction) (*e*) the Order is balanced without interaction assumptions but analyzed by a modified latin square design (e.g., Lindquist Type II design); and (*f*) the Order is balanced without assumptions but analyzed as a simple Treatments×Subjects design.

Repeated Measurements Designs

Case I. Same Order

This represents the simplest type of repeated measurements design. All *n* *S*s receive the *k* Treatments in the same Order. Table 1 indicates the E(*MS*). The rule for obtaining the E(*MS*) in a complete factorial design is as follows: E(*MS*) is σ_e^2 (variance due to error) plus the σ^2 term whose subscript corresponds to the main or interaction effect of concern. It also includes all σ^2 terms which represent interactions with this main or interaction effect, providing the effects not included in the main or interaction effect are all random. For example, in a two-variable design ($A \times B$) in which *A* is a random effect, the E(*MS*) for *B* would be $\sigma_e^2+\sigma_b^2+\sigma_{ab}^2$; for *A*, $\sigma_e^2+\sigma_a^2$ (see Anderson & Bancroft, 1952; Cornfield & Tukey, 1956; Greenwood, 1956; Kempthorne, 1952; Wilk & Kempthorne, 1957). The coefficient for σ_e^2 is 1; all other σ^2 terms have coefficients which are equal to the number of replications (*n*) times the number of the levels of the variables which are not included in the subscript of the σ^2 under consideration.[2] However, because of confounding aspects other components will be included in some mean squares. These can be determined intuitively.

TABLE 1

COMPONENTS OF VARIANCE INCLUDED IN MEAN SQUARES FOR CASE I: ORDER EFFECT PRESENT

T	$\sigma_e^2+s\sigma_t^2+\sigma_{ts}^2+s\sigma_o^2$
S	$\sigma_e^2+t\sigma_s^2$
TS	$\sigma_e^2+\sigma_{ts}^2$

Note.—In Tables 1–6 all main effects are fixed except *S*, which is random.

[2] The most general treatment of coefficients for a complete factorial design is by Cornfield and Tukey (1956). They use $1-x/X$ as coefficients for σ_e^2 where *x* refers to the number of replications and *X* is the total population. $1-x/X$ is also used as a coefficient for *x* in σ^2 due to interactions if *x* is not involved in the mean square in question. These coefficients serve to suppress terms completely when they are fixed effects (if $x=X$, then coefficient is zero). If *x* is very small and *X* is infinite or very large, then the coefficient goes to 1. These coefficients also reduce the σ^2 terms (are between zero and 1) when the populations are finite but larger than the samples. We shall not be concerned with this latter situation. Thus in Table 1 the coefficient for σ_e^2 throughout the table and for σ_{ts}^2 in *T* is $1-s/S$. Inasmuch as the sample of *S*s is small whereas the population *S* is usually very large, the coefficient becomes 1. The coefficient for σ_{ts}^2 in *S* is $1-t/T$; however, $t=T$ because we have all Treatment levels in our experiment. Thus the coefficient is zero and σ_{ts}^2 vanishes. The coefficients for σ_{ts}^2 in *TS* is 1 because $1-t/T$ does not appear for either the *T* or *S* portions. Both are involved in the *TS* mean square. The coefficient for σ_{ts}^2 in *T* is $1-s/S$, which becomes 1. All coefficients are multiplied by *n*, which in our example is 1.

In Table 1 Subjects represents a random variate and thus σ_{ts}^2 appears in the Treatments effect. No Order ×Treatment effect, or any interaction containing Order, is present because only one order is involved.[3] Furthermore, σ_o^2 and σ_t^2 are confounded and cannot be separated. The coefficient for σ_o^2 is the same as for σ_t^2. Of course, if $\sigma_o^2=0$ then the test of T by TS is a valid one. If σ_o^2 is not zero then positive bias occurs in the F test of T, a tendency for too many significant effects being reported (Type I error).

Case II. Randomization of Order

Because we suspect the presence of an Order effect (and interactions involving this effect) we decide to randomize the Order of presentation of the Treatments to each individual separately. The result of this procedure is indicated by Table 2. In this case σ_o^2 has been removed from T. Any effects of Order or any interaction will appear in σ_e^2 and, thus, be felt in all effects, so that the F test of T will be a valid one. However, σ_e^2 will be inflated if the Order or interaction effects are present.

TABLE 2

COMPONENTS OF VARIANCE INCLUDED IN MEAN SQUARES FOR CASE II: RANDOMIZATION OF ORDER

T	$\sigma_e^2+s\sigma_t^2+\sigma_{ts}^2$
S	$\sigma_e^2+t\sigma_s^2$
TS	$\sigma_e^2+\sigma_{ts}^2$

Case III. Balanced Design—Interactions Zero

This situation represents the limited example which has been considered in detail by Gaito (1958a). Hilgard (1951) and Lindquist (1953), as well as others, have been concerned with this case. Because of possible practice effects we have one or more *S*s take one Order, one or more other *S*s take another Order, etc., but assume that all Order interactions are zero. Table 3 indicates that σ_o^2 now appears in TS, thus making for negative bias in the F test of T. The balancing procedure equalizes the various levels of each main effect but not the interaction components. If more than one factor is included in the experiment, the levels of all main effects and interactions not involving the random effect are equalized, but the interaction levels including the random effect are not. Furthermore, the magnitude of σ_o^2 inflation tends to increase with increasing order of interaction (e.g., mean square of T_1T_2S is greater than T_1S or T_2S).

Case IV. Single Latin Square—Interactions Present

The single latin square design has been used infrequently in recent years in psychology, possibly because of the criticisms of Lindquist (1953) and of McNemar (1951) concerning the frequent presence of interactions. This design (in which each S has a different Order) has been considered by Gaito (1958b) as the One Random

TABLE 3

COMPONENTS OF VARIANCE INCLUDED IN MEAN SQUARES FOR CASE III: BALANCING OF ORDER BUT NO ORDER INTERACTIONS PRESENT

T	$\sigma_e^2+s\sigma_t^2+\sigma_{ts}^2$
S	$\sigma_e^2+t\sigma_s^2$
TS	$\sigma_e^2+\sigma_{ts}^2+s\sigma_o^2$

[3] Even though we speak of Order effects and Order interactions, it is actually trial effects and trial interactions which are involved, because differences between trials, or differential effects of trials for different *S*s, indicate that the different Orders do not give the same results. However, it is usual to speak in terms of the former.

TABLE 4

Components of Variance Included in Mean Squares for Case IV: Single Latin Square Analysis with No Assumption Concerning Interactions

T	$\sigma_e^2+t\sigma_t^2+\sigma_{so}^2+\sigma_{st}^2+(1-2/t)\sigma_{sto}^2$
O	$\sigma_e^2+t\sigma_o^2+\sigma_{st}^2+\sigma_{so}^2+(1-2/t)\sigma_{sto}^2$
S	$\sigma_e^2+t\sigma_s^2+\sigma_{to}^2+(1-1/t)\sigma_{sto}^2$
Residual	$\sigma_e^2+\sigma_{to}^2+\sigma_{st}^2+\sigma_{so}^2+(1-2/t)\sigma_{sto}^2$

Note.—Because $o=t=s$, the coefficients for all main effects are given as t.

Variate Model. That article also deals with the Zero, Two, and Three Random Variates Models as well. The rule for the complete factorial design presented above must be modified to deal with this incomplete factorial design. The above rule is applied first. Then the following additions are included. Residual contains all interactions, and each main effect is confounded with the triple interaction and the double interaction containing the other two effects. The paper by Wilk and Kempthorne (1957) presents a generalized derivation for latin square designs and the coefficients for each σ^2 term in Table 4 are based on that derivation. σ_e^2 and all interaction σ^2 terms except σ_{sto}^2 have a coefficient of 1. The coefficient of σ^2 for each main effect is t. The coefficient for σ_{sto}^2 in the Residual and the two fixed effects (T and O) is $1-2/t$; in S the coefficient gets closer to one $(1-1/t)$ inasmuch as the random effect is involved.

In this case the F test of T is negatively biased unless σ_{to}^2 is zero. Even though the F test is unbiased when σ_{to}^2 is zero it is not a valid F test because it is not distributed as the F distribution. A valid F test requires that the interactions in the mean squares of both the main effect and the Residual must be random, normally distributed, and be a component that would be expected in the mean square as indicated by the rule above. If these conditions are not satisfied the result is a ratio of two noncentral chi square statistics divided by their respective degrees of freedom, and the distribution depends upon the parameters of unwanted components, in the present situation σ_{so}^2 and σ_{sto}^2. For a valid and unbiased F test, σ_{so}^2, σ_{to}^2, and σ_{sto}^2 must be zero.

Case V. Lindquist Type II Design—Interactions Present

This situation is the same as in Cases III and IV except that groups of *S*s take each Order; also we allow all Order interactions to be present and analyze the results as a Lindquist Type II design (Table 5). This design is actually a modification of the single latin square design and the arguments presented above for the One Random Variate Model are pertinent here. The Residual contains σ_e^2 and all possible interactions except σ_{to}^2, which has been removed. Each main and interaction effect contains σ_e^2, variance due to itself, and the interaction of the effect with other effects which are random. Furthermore, because of the confounding aspects of the latin square each main effect includes variance due to the other two factors and variance due to the triple interaction. In this design if only two Treatments and two Orders are involved the $T\times O(w)$ effect disappears (Lindquist, 1953).

The F test of T and $T\times O(b)$ will

TABLE 5

Components of Variance Included in Mean Squares for Case V: Balancing of Order, All Interactions Present, and Analyzed as Lindquist Type II Design

Between *S*s	
Groups or $T \times O(b)$	$\sigma_e^2 + t\sigma_s^2 + s't\sigma_{to}^2(b) + (1-1/t)\sigma_{sto}^2$
Between *S*s within Groups	$\sigma_e^2 + t\sigma_s^2 + (1-1/t)\sigma_{sto}^2$
Within *S*s	
Treatments	$\sigma_e^2 + s\sigma_t^2 + \sigma_{ts}^2 + \sigma_{so}^2 + (1-2/t)\sigma_{sto}^2$
Order	$\sigma_e^2 + t\sigma_o^2 + \sigma_{so}^2 + \sigma_{ts}^2 + (1-2/t)\sigma_{sto}^2$
$T \times O(w)$	$\sigma_e^2 + s'\sigma_{to}^2(w) + (1-2/t)\sigma_{sto}^2$
Residual	$\sigma_e^2 + \sigma_{ts}^2 + \sigma_{so}^2 + (1-2/t)\sigma_{sto}^2$

Note.—Because $o = t$, t is used as the coefficient for both σ_s^2 and σ_o^2. s' refers to the number of *S*s in each group, s to the number of *S*s in the experiment.

not be biased even if all interactions are present. However, the *F* test of $T \times O(w)$ will be negatively biased. The unbiased tests will not be distributed as *F* because of nuisance parameters, e.g., in Treatments, the σ_{so}^2 and σ_{sto}^2. Thus the Type II "mixed" design, which is one of a number of designs which Lindquist recommends for counterbalancing purposes (1953, p. 163, Ch. 13) appears to give unbiased (but nonvalid) results for the main effects, when all interactions are present.

The advantage of the Type II design is that it allows for a separation of both the Order and the Treatments ×Order effects. However, if the latter is present the test of the Treatments effect may not be meaningful, even though unbiased. If Order were a random effect, then the test of the Treatments effect is meaningful. However, usually Order represents a fixed effect. Thus if the interaction is of a "reversal" type (i.e., one Treatment is most effective with one Order of presentation whereas other Treatments are more effective with different Orders), an *F* test of *T* would be meaningless. However, in a "continuous spread" type of interaction (i.e., the rank order of the Treatments are the same for all Orders but the difference between Treatments varies with the Order of presentation), a generalization based on the *F* test of *T* would be meaningful.

The Type II design represents one of a large number of "mixed" designs. Readers interested in the E(*MS*) for more of these should consult Harter and Lum (1955).

Case VI. Balanced Treatments×Subjects Design—Interactions Present

Let us take the same procedure as in Case V but analyze the results as a simple Treatments×Subjects design. This result is indicated in Table 6. The E(*MS*) for *T* is the same as in Table 5; *S* contains all the between-subjects variance terms of that table; and *TS* contains the *O*, $T \times O(w)$, and Residual components. Note that σ_o^2 is contained in *TS* as was indicated for Case III. The reader should note also that *TS* is the same in Tables 3 and 6, except that in the latter table are included Order interaction σ^2 terms while in Table 3 these are missing. For the E(*MS*) of Table 3 it was assumed that Order interactions were not present.

As is obvious from Table 6, the *F* test of *T* will be negatively biased because two unwanted components,

TABLE 6

Components of Variance Included in Mean Squares for Case VI: Balancing of Order and All Order Interactions Present

T	$\sigma_e^2+s\sigma_t^2+\sigma_{ts}^2+\sigma_{so}^2+(1-2/t)\sigma_{sto}^2$
S	$\sigma_e^2+t\sigma_s^2+s'l\sigma_{to}^2(b)+(1-1/t)\sigma_{sto}^2$
TS	$\sigma_e^2+t\sigma_o^2+\sigma_{ts}^2+\sigma_{so}^2+s'\sigma_{to}^2(w)+(1-2/t)\sigma_{sto}^2$

σ_o^2 and $\sigma_{to\ (w)}$, will be included in the denominator. The defects occurring in this situation are more severe than in the above cases.

Discussion

From the six cases presented above it is obvious that the possible defects which may occur in repeated measurements designs are extreme. It would appear that if one does have a repeated measurements design, the safest procedure would be to randomize the order of treatments so that order and all interactions containing order would be included in σ_e^2 and appear in all effects, unless he has strong reasons for believing that certain interactions are not present. However, one might use a Lindquist Type II design. In the former design unbiased and valid F tests of the treatment effect are obtained. In the latter design unbiased tests of the treatment effect are obtained but these tests are not distributed as the F distribution and will not be meaningful if a "reversal" type interaction between order and treatments has occurred.

All of these counterbalancing examples have been of an intersubject nature. However, the results of intrasubject counterbalancing would be similar. For example, if intrasubject counterbalancing were used such that each S would receive two or more sequential orders (e.g., if two treatments, the *S*s would take only the ABBA or BAAB orders), the σ_o^2 would be confounded with σ_t^2 in the treatments effect. If inter- and intrasubject counterbalancing were to be employed, some *S*s would receive two or more orders of presentation while other *S*s would receive some reversal of these orders (e.g., if two treatments, some *S*s would have the ABBA sequence while others would have the BAAB sequence). In this case if a subjects×treatments analysis is followed and a practice effect is present which is constant from trial to trial for all *S*s, no bias occurs in either the main effects or the interaction; however, the within-cells term will be inflated. If the practice effect is not constant from trial to trial, and is either the same or not for all *S*s, then inflation will occur in both the interaction and within-cells terms.

The above considerations should make one cautious concerning the use of a repeated measurements design. However, only the effects of order and interactions have been discussed. There is another source of contamination in the repeated measurements designs, i.e., correlated observations. It has been assumed by many investigators that by partitioning a source of variation attributable to *S*s, the problem of correlation has been handled. That this assumption is not true has been indicated by a number of people (e.g., Box, 1954; Danford & Hughes, 1957; Geisser & Greenhouse, 1958; Lubin, 1954, 1957, 1958; Scheffe, 1956).

Box (1954) indicates that when there is moderate correlation within rows (in psychological experiments the row variable would represent *S*s), a great distortion occurs in the probability levels for between-rows comparisons but little distortion is introduced for between-columns comparisons. The maximum correlation that Box studied was ±.40. In the case of the negative correlation the percent probability for the test of columns (Treatments) was 5.90 rather than 5.00 (which would result when correlation is zero); for positive correlation the percent probability was 6.68. Box makes use of an approximate technique in which the degrees of freedom are reduced by multiplying each *df* by a fraction, epsilon (ϵ), which depends on the correlation within-rows. The upper limit of ϵ is 1, which will occur only if the variances are equal and the correlation is constant among the Treatments. In this case the *F* ratio with the usual *df* can be used. In the event that just two treatments are involved, ϵ equals 1 if the variances are equal. However, in many designs using three or more treatments, ϵ will be less than 1; thus if the usual *df* are employed (without reduction by ϵ) an increase in Type I errors will occur.

Geisser and Greenhouse (1958) have extended Box's result to develop a conservative *F* test of treatments. They show that $\epsilon \geq (k-1)^{-1}$ and thereby determine the lower limit for the *df* to be $1/n-1$, where k refers to the number of treatments and n is the number of *S*s. (This result can be obtained by multiplying the *df* for treatments $(k-1)$ and for treatments×subjects $[(k-1)(n-1)]$ each by $1/k-1$.) Thus the *F* test with *df* of 1 and $n-1$ can be employed when unequal covariation occurs with one group of *S*s. They also develop a conservative test when more than one group is involved. In this case the *df* for the approximate *F* test of treatments is $1/N-g$, where N is the total number of *S*s and g is the number of groups. However, the authors maintain that the use of the lower limit may be too conservative.

Danford and Hughes (1957) argue for the use of the usual analysis of variance design, maintaining that the equal covariance assumption (constant correlation) is tenable for certain experimental situations. They state that some experimental data have shown comparable correlation coefficients (*r*'s of .70 to .90).[4] They criticize Scheffe's (1956) suggestion to use Hotelling's T^2 statistic for testing the fixed main effect because of the above. Likewise, they indicate that if the equal covariance assumption is correct the power of the usual *F* test is greater (in some cases, much greater) than is the power of Hotelling's test.

Lubin (1954, 1957, 1958) has cogently considered the repeated measurements designs, not only considering the effects of correlated observations but also treatment×order interactions, and other learning or "carry over" effects. Because of these contaminating effects, he recommends the use of a modification of Hotelling's T^2 test, or a nonparametric rank-order test if one is interested in the relative efficacy of several treatments (unless a treatment ×order interaction is present). If this interaction is present he advocates a matched *S*s design in which each *S* receives only one treatment.

Thus the *F* test is theoretically correct only if constant correlation among treatments is present. If only two treatments are involved, and homogeneity of variance is present, then it follows that the *F* test is always appropriate. If unequal corre-

[4] The experimental data of concern are not cited, however.

lation occurs, too many significant *F*s will be reported. With moderate, but unequal, correlation among treatments, the increase in number of significant results reported for treatments effects appears to be small, using Box's approximation. The increase with greater correlation is unknown. However, the *F* test indicated by Geisser and Greenhouse allows one to make a conservative test.

In conclusion, it is apparent that multiple defects are present in repeated measurements designs. The design using randomization of the order of treatments avoids the numerous defects but σ_e^2 may be quite large. With randomization the correlation effect should be minimized. The Lindquist "mixed" design overcomes some of the defects but the *F* test of treatments, even though unbiased may not be a valid *F* test and may be meaningless. The matched *S*s design recommended by Lubin would appear to be the safest procedure if enough *S*s are available. However, the important point to stress is that if an investigator resorts to a repeated measurements design he should be aware of possible distortions which may occur and be able to defend his assumptions concerning the order effect, the order interactions, and the correlated observations.

Summary

Six types of analysis of repeated measurements designs are indicated. The effects of order, interactions containing order, and correlated observations on the components of variance and analysis of variance tests of significance are considered. The first two act, in general, to inflate the error estimates and thus to increase the probability of a Type II error. The correlated observations (if unequal) have the opposite effect, i.e., increase the probability of a Type I error. It is suggested that caution be exercised in the use of these designs; randomization of the order of treatments or matched subjects appear to be the safest procedures. The Lindquist Type II "mixed" design overcomes some defects but is not completely appropriate.

REFERENCES

Alexander, H. W. The estimation of reliability when several trials are available. *Psychometrika*, 1947, **12**, 79–99.

Anderson, R. L., & Bancroft, T. A. *Statistical theory in research.* New York: McGraw-Hill, 1952.

Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems: II. Effects of inequality of variance and correlation between errors in the two way classification. *Ann. math. Statist.*, 1954, **25**, 484–498.

Cornfield, J., & Tukey, J. W. Average values of mean squares in factorials. *Ann. math. Statist.*, 1956, **27**, 907–949.

Danford, M. B., & Hughes, H. H. Mixed model analysis of variance, assuming equal variances and equal covariances. *USAF Sch. Aviat. Med. Rep.*, 1957, No. 57-144.

Gaito, J. Statistical dangers involved in counterbalancing. *Psychol. Rep.*, 1958, **4**, 463–468. (a)

Gaito, J. The single latin square design in psychological research. *Psychometrika*, 1958, **23**, 369–378. (b)

Garrett, H. E., & Zubin, J. The analysis of variance in psychological research. *Psychol. Bull.*, 1943, **40**, 233–267.

Geisser, S., & Greenhouse, S. W. An extension of Box's results on the use of the *F* distribution in multivariate analysis. *Ann. math. Statist.*, 1958, **29**, 885–891.

Gourlay, N. *F*-test bias for experimental designs of the latin square type. *Psychometrika*, 1955, **20**, 273–287.

Grant, D. A. On "The analysis of variance in psychological research." *Psychol. Bull.*, 1944, **41**, 158–166.

Grant, D. A. The latin square principle in the design and analysis of psychological experiments. *Psychol. Bull.*, 1948, **45**, 427–442.

GREENWOOD, J. A. Analysis of variance and components of variance: Factorial experiments. Unpublished paper, USN Bureau of Aeronautics, 1956.

HARTER, H. L., & LUM, M. D. Partially hierarchal models in the analysis of variance. *USAF WADC Rep.*, 1955, No. 55-33.

HILGARD, E. R. Methods and procedures in the study of learning. In S. S. Stevens (Ed.), *Handbook of experimental psychology*. New York: Wiley, 1951. Pp. 517–567.

KEMPTHORNE, O. *The design and analysis of experiments*. New York: Wiley, 1952.

KOGAN, L. S. Analysis of variance: Repeated measurements. *Psychol. Bull.*, 1948, **45**, 131–143.

KOGAN, L. S. Variance designs in psychological research. *Psychol. Bull.*, 1953, **50**, 1–40.

LINDQUIST, E. F. Goodness of fit of trend curves and significance of trend differences. *Psychometrika*, 1947, **12**, 65–78.

LINDQUIST, E. F. *Design and analysis of experiments in psychology and education*. New York: Houghton Mifflin, 1953.

LUBIN, A. Are non-parametric tests distribution-free? Unpublished paper, Walter Reed Army Institute of Research, 1954.

LUBIN, A. Some rank-order tests for trend in a set of correlated means. Unpublished paper, Walter Reed Army Institute of Research, 1957.

LUBIN, A. On the repeated measurements design. Unpublished paper, Walter Reed Army Institute of Research, 1958.

MCNEMAR, Q. On the use of latin squares in psychology. *Psychol. Bull.*, 1951, **48**, 398–401.

MCNEMAR, Q. *Psychological statistics*. New York: Wiley, 1955.

PETERS, C. C. Interaction in analysis of variance interpreted as intercorrelation. *Psychol. Bull.*, 1944, **41**, 287–299.

SCHEFFE, H. A mixed model for the analysis of variance. *Ann. math. Statist.*, 1956, **27**, 23–36.

WILK, M. B., & KEMPTHORNE, O. Non-additivities in a latin square design. *J. Amer. Statist. Ass.*, 1957, **52**, 218–236.

(Received August 6, 1959)

Psychological Bulletin
1961, Vol. 58, No. 1, 55–79

HUMAN TRACKING BEHAVIOR[1]

JACK A. ADAMS
University of Illinois

The subject matter of this paper is a critical review and analysis of research, issues, and points of views associated with human behavior in one- and two-dimensional tracking tasks. Tracking tasks have never been given explicit definition and one of the purposes of this paper is to tentatively advance the general bounds of a tracking situation, but for those who are unfamiliar, a temporary working definition for the moment which enjoys the consensus of most psychologists is as follows:

1. A paced (i.e., time function) externally programed input or command signal defines a motor response for the operator, which he performs by manipulating a control mechanism.
2. The control mechanism generates an output signal.
3. The input signal minus the output signal is the tracking error quantity and the operator's requirement is to null this error. The mode of presenting the error to the operator depends upon the particular configuration of the tracking task but, whatever the mode, the fundamental requirement of error nulling always prevails. The measure of operator proficiency ordinarily is some function of the time-based error quantity.

The usual tracking task has a visual display but there is no necessity for this. On occasion, auditory tracking tasks have been devised (Forbes, 1946; Humphrey & Thompson, 1952a, 1952b, 1953). The most simple and well-known visual tracking task is the Rotary Pursuit Test (Melton, 1947) which employs a repetitive input signal and, although investigations using the Rotary Pursuit Test are not ordinarily included in that body of research which is considered to study tracking behavior per se, it is nevertheless an unequivocal example of the breed. Tracking studies typically use more elaborate apparatus which allows for controlled manipulation of such variables as the function for the input signal, scale factors, mathematical transformations of the output signal, characteristics of the control mechanism, etc.

While investigations of tracking behavior might legitimately be subsumed under the time-honored rubric "motor skills," this label is misleading in hinting by implication and textbook tradition that motor behavior, such as tracking, is disassociated from so-called "higher processes." British investigators in particular have analyzed the acquired ability to predict input stimulus sequences as a key intervening response class in determining the proficiency of the measured motor responses in tracking tasks, thus emphasizing the interlacing of "higher" and "lower" processes. These British studies will be discussed in detail later, but passing mention of them at the onset seems worthwhile for establishing the archaic connotations of "motor skills." Research by Adams (1957), Fleishman (1954, 1957a, 1957b, 1958), and Fleishman and Hempel (1954, 1955, 1956) on variables influencing individual differences in motor behavior, also docu-

[1] This research was supported by the United States Air Force under Contract No. AF 49(638)-371, monitored by the Air Force Office of Scientific Research of the Air Research and Development Command.

A number of psychologists read this manuscript in draft form and contributed to its improvement. The detailed and critical comments of F. C. Bartlett, E. A. Fleishman, C. B. Gibbs, N. B. Gordon, J. A. Leonard, and A. T. Welford were particularly appreciated.

ments the inherent complexity of the response totality elicited in motor tasks. Helson (1949), in discussing variables influencing the subject's standard of excellence in a tracking task, includes perceptual and motivational states in addition to motor factors as influential determiners of motor behavior.

Basic Terminology and Frame of Reference

Independent variables influencing tracking behavior will be divided into two classes: task variables and procedural variables. *Task variables* are machine-centered. They are the physical values of the tracking device, and they include such factors as the nature of the input signal, configuration of the display, design of the control system, mathematical transformations relating control displacement and changes in the output signal, etc. *Procedural variables* are man-centered. They are manipulable nontask quantities, and examples of them are instructions, number of practice trials, length of the practice trial, and time between trials. Also, the indicants which are displayed to the operator will be implicitly assumed as simple elements, such as needles or dials, pointers, dots on cathode ray tubes, etc. Special problems that arise when the display is perceptually complex and requires the interpretation of forms, shapes, colors, etc. will be ignored.

The Tradition of Engineering Psychology

A dominant influence in tracking research is the experiments of engineering psychology, with the emphasis being largely on the relations between measures of tracking behavior and *task* variables. The engineering psychologist has as his goal the prediction of the characteristics of man-machine systems, and this goal requires careful attention to the task variables which influence the operator. Representative examples of several hundred task-oriented tracking experiments are studies of control loadings (Bahrick, 1957; Bahrick, Bennett, & Fitts, 1955; Bahrick, Fitts, & Schneider, 1955; Briggs, Bahrick, & Fitts, 1957; Howland & Noble, 1953; Weiss, 1954), input signal characteristics (Hartman, 1957; Hartman & Fitts, 1955; Noble, Fitts, & Warren, 1955), the magnitude of lag between control movement and system output (Conklin, 1957; Warrick, 1949), the effects of visual noise (Briggs & Fitts, 1956; Briggs, Fitts, & Bahrick, 1957), mathematical transformations of the output signal (Briggs, Fitts, & Bahrick, 1958), and compensatory vs. pursuit tracking (Chernikoff & Taylor, 1957; Poulton, 1952b). Task variables, because of their role in determining the behavioral requirements for the operator, are an important class of variables for psychology and engineering psychologists have made a notable contribution in directing attention toward neglected determiners of human behavior. However, this strong task orientation has led to the neglect of procedural variables that influence the operator, and thus the efficiency of the total man-machine system. A recent article (Taylor, 1957) has clearly stated this emphasis:

> . . . human engineering aims first at building better systems and secondarily at improving the lot of the operator. Thus, whereas conventional psychology, both basic and applied, is anthropocentric, human engineering is mechanocentric (p. 252).

This statement succinctly summarized the task-oriented approach of engineering psychology and expresses a downgrading of procedural variables related to training, retention, fatigue, motivation, etc. It is forgotten, or intentionally neglected, that the engineering psychologist

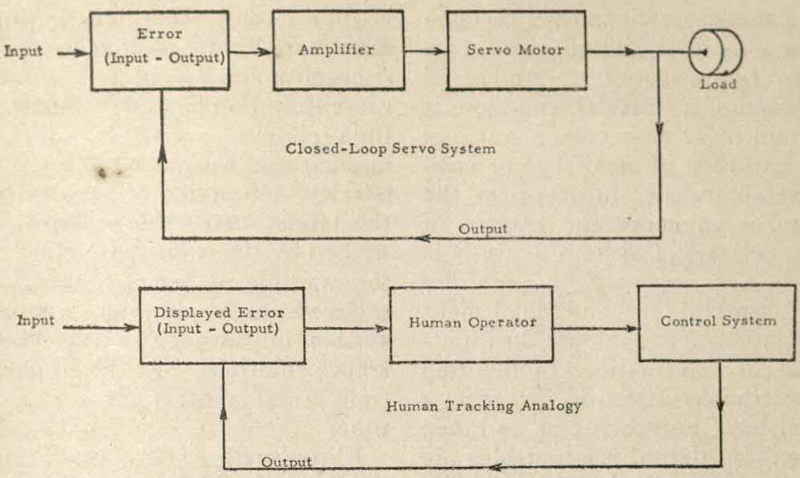


FIG. 1. The analogy commonly drawn between a closed-loop electromechanical servosystem and a human operator as an error-nulling agent in a tracking task.

must, over the long run, develop the capability to predict the effectiveness of a man-machine system for different states of the operator, and this means a strict scientific accounting of a broad range of variables which influence man. There are few who underestimate the importance of task variables in determining the behavior of a man-machine system, but there seems to be no sound justification for relegating procedural variables to a secondary status. In the beginning, an applied branch of a science might profitably concern itself with rank ordering its variables in terms of their potency in influencing a criterion (Taylor & Garvey, 1959), but this approach does not deserve being elevated to a research philosophy. Sophisticated applied science, just as sophisticated basic science, must work toward a precise accounting of all variables and their interrelations.

Whereas general experimental psychology has often looked to traditional behavioral theory as a basis for its tracking studies, many engineering psychologists, with their mechanocentric views, have turned towards the feedback theory of closed-loop servomechanisms (Bower & Schultheiss, 1958; Brown & Campbell, 1948; Goode & Machol, 1957) as a model for a man-machine tracking system. Basically, a closed-loop servosystem is an electromechanical error-nulling system which compares an input signal with an output signal and works toward reducing the difference between them. Because error nulling is a basic characteristic of systems which include the human operator as a tracking component, some engineering psychologists view physical servotheory as a potential source of descriptive relationships for manual tracking systems. Figure 1 shows the parallel that is ordinarily drawn between a servosystem and a man-machine tracking system. The theory of servomechanisms is a method of mathematical analysis concerned with the description of the output of a complex system as a function of the input, and it allows the system's analyst to state the functional characteristics of his system with some precision. The expression of the input-output relations is by means of a complex ratio called the transfer function which expresses the nature of the transformations that the system imposes on the input signal. In a system comprised of a number of components, a transfer function is determined for each component and these can then be combined to yield an overall transfer function for the system. An important feature of these methods of system analysis is that is not necessary to painfully trace the signal through each element of a component to compute the input-output transformation represented in the transfer function. Rather, a "black box" approach can be taken where input-output relationships are directly compared without attending to the many intermediate transformations which occur to the signal as it passes through the component.

The servosystem analyst is concerned with input-output relations as they are manifested in two domains: time and frequency. In the time domain the time-varying characteristics of the system are described in terms of overshooting, undershooting, oscillations, steady state errors, etc., in response to a step input. In the frequency domain the output of the system is examined for transformations of a sinusoidal input after the transients have died out. Finally, and perhaps most importantly in this brief exposition on the methods of servosystem analysis, is that the entire mathematical structure is founded on the assumption of linearity. Fundamentally, this assumption means that the system obeys the superposition theorem which states that the system response to the sum of a set of inputs is equal to the sum of the responses made to each input separately. This means that the performance of the system can be predicted for any complex input providing we know the response of the system to each of the constituent inputs comprising the complex input. Another implication for the linear assumption is that it will accurately reproduce input sinusoidal frequencies after transients have died out, although there may be phase shift and amplitude change. Furthermore, it is implicit that the output of the system is solely a function of the input and this functional relationship is described by the transfer function—i.e., for example, it is not a function of such variables as time where the system might perform one class of transformations on the inputs at time *t* and another class at a later time.

Ellson's paper (1949) best expresses the hope of some engineering psychologists that the transfer function for the human operator might be determined and provide an analytical means of predicting the performance of the total man-machine system, and of optimizing the performance of the system by designing hardware components to complement the response characteristics of man. This goal of mathematically describing the characteristics of man and his machine elements is scientifically admirable but, regrettably, it was doomed from the beginning by the massive barrier of the linearity assumption. Almost self-evident is the fact that the human operator is a nonlinear component of a system with his intricate adaptive propensities toward learning, fatiguing, motivational shifts, etc. and that there is faint possibility of finding *the* transfer function which can be used by system designers to optimize the performance of a system by capitalizing on the transformations that man imposes on a signal as it enters the receptors, makes passage through the organism, and is emitted anew by the responding effector system (Birmingham & Taylor, 1954; Ellson, 1949; Fitts, 1951; Searle & Taylor, 1948). Birmingham and Taylor (1954) have nicely expressed this matter of nonlinearity for the tracking human operator:

> This adaptability on the part of the man is, of course, a great boon to the control designer, since he can rely upon the human to make the most of any control system, no matter how inadequate. It is this which probably constitutes the most important single reason for using men in control loops. Yet, this very adjustability renders any specific mathematical expression describing human behavior in one particular control loop quite invalid for another man-machine arrangement. This suggests strongly that "*the* human transfer function" is a scientific will-o'-the-wisp which can lure the control system designer into a fruitless and interminable quest (p. 1752).

Fitts (1951) has reported on certain limited conditions where human response appears to approximate linearity but, in general, it would seem that the nonlinearities of human behavior negate the usefulness of the servomodel and its mathematical techniques as a serious theoretical instrument for behavior theory or as a tool for the design of man-machine systems. Nonlinearities do, of course, occur in some physical systems but the assumption of linearity is met sufficiently well and often to make the theory of important value for the physical sciences. This could hardly be said for psychology where nonlinearities are an inherent, and indeed the most interesting and challenging, aspect of the human operator. It must be concluded therefore, that present-day servotheory stands in an analogous, not a scientific, relationship to man-machine tracking systems.

Even if analytical methods eventually become available to handle the nonlinearities of closed-loop human behavior, it is unlikely that engineering psychology will be able to make effective use of them if it continues its preoccupation with task variables (Taylor, 1957; Taylor & Garvey, 1959) and underplays the role of procedural variables which are basic determiners of dispositional states of the operator and contribute substantially to the nonlinearities. Engineering texts on servotheory (Bower & Schultheiss, 1958; Brown & Campbell, 1948; Goode & Machol, 1957) distinguish between *analysis* or the description of a system of existence, and *synthesis* or the prediction of the characteristics of components of the system to achieve certain objectives. Conceivably, we might eventually describe a man-machine tracking system already in existence because the response characteristics of the human operator can be empirically determined for the range of inputs of interest and the operator states that prevail. However, synthesizing is quite different because it requires that we know the laws of human behavior as a function of task and procedural variables and are able to *predict* the characteristics of the human response functions. Questions relating to such operator states as learning and fatigue most certainly will arise and it is evident that these queries will not be answerable if task variables are taken as the primary research domain of engineering psychology. Engineering psychology, it would seem, cannot escape the burden of the same variables and searches for lawfulness which traditionally occupy all psychologists.

In defense of the servotheory approach to tracking, its protagonists have been engaged in proper search for a descriptive mathematical device for man-machine tracking systems which includes provisions for task variables and the properties of response outputs to inputs which are continuous with respect to time. Contemporary behavior theories ordinarily employ measures of behavior, such as frequency and latency, which can be defended as operationally meaningful dependent variables but which are gross summary indices of complex behavior sequences and often do violence to the subtleties of the ongoing behavior. Commonly, psychologists in their laboratory research will elicit elaborate time-based response sequences from an organism and then will ignore completely the time-varying characteristics of the responding in their measurement. In contrast, psychologists studying tracking have recognized, almost from the beginning, that their dependent measures should somehow describe the prominent characteristics of time-based response functions. And, because contemporary behavior theories give no attention to time functions, tracking psychologists appear to have suffered disenchantment and have turned to the mathematical schema of closed-loop servotheory, inadequate though it is, because it grapples directly with the measurement and description of time-varying quantities. The fact that servotheory is of little value for quantitative description of man-machine tracking systems should not allow us to forget that the interest in it has reflected a legitimate concern about measurement issues and variables which are important for the response phenomena under investigation.

Tradition of General Experimental Psychology

Basic research on tracking by general experimental psychologists has not had the strong emphasis of task variables. Frequently, in basic research, the experimental task has been a convenient means of eliciting a response class for the purposes of manifesting underlying behavioral processes which are of theoretical rather than practical interest, and consequently tracking tasks have not been studied for their own sake. Examples of this approach are many of the tracking studies on the Rotary Pursuit Test with interest in fatigue-like effects or, more exactly, the implications of Hull's (1943) expressions of reactive and conditioned inhibition for behavior (Adams, 1956; Adams & Reynolds, 1954; Kimble & Horenstein, 1948). Other studies of fatigue processes (Floyd & Welford, 1953; Payne & Hauty, 1954; Siddall & Anderson, 1955) using tracking tasks have had a similar general concern and have shown little interest in the study of tracking for its own sake. The interest in task variables per se which has preoccupied engineering psychology has been largely absent in the research of general experimental psychology. This has been a healthy countertrend to the task emphasis of engineering psychology but the approach of using virtually any convenient task to elicit a response class can be considered a deficiency because it shows a lack of appreciation for the influence of task variables on behavior, and the possible interactions that can be expected to occur between task and procedural variables. These studies seem to have implicitly assumed that behavioral laws will transcend particular characteristics of a task, but this is an unlikely possibility because of the extensive work in engineering psychology showing the potent influence of task variables on performance. There is good reason to expect that many task variables will interact with those variables which have been of interest in testing theoretical deductions. To illustrate, if it were eventually found that a major cause for the depressant effects of massed practice on the tracking response was that work inhibition degraded the quality of proprioceptive feedback, the behavior functions would, as a minimum, have to be expressed in relation to the interaction of intertrial interval and those control system variables which determine proprioceptive feedback. Helson (1949), in a report of the Foxboro investigations which were an early series of systematic tracking studies, points out that both task and procedural variables are pertinent to a complete understanding of human behavior. Lewis (1953) has urged closer attention to the relations between the physical organization of tasks and the complexities of behavior.

An important line of tracking research, which can be subsumed under the rubric of general experimental psychology, has been dominated by British investigators of the Applied Psychology Research Unit, Cambridge University, and mainly concerns efforts to delineate the intrinsic characteristics of the overt motor tracking response, and to identify and assess the response classes which intervene between the displayed stimuli and the measured motor response. Examples of these interests are the question of whether the apparently smooth, continuous tracking response is fundamentally intermittent (Chernikoff & Taylor, 1952; Craik, 1947, 1948; Davis, 1956; Elithorn & Lawrence, 1955; Hick, 1948; Poulton, 1950; Searle & Taylor, 1948; Taylor & Birmingham, 1948; Vince, 1948a, 1948b, 1949; Welford, 1952) and the conditions under which the human operator learns to predict or anticipate changes in the input signal (Bartlett, 1951; Craik, 1947, 1948; Leonard, 1953; Poulton, 1952a, 1957a, 1957b, 1957c; Vince, 1953, 1955). These studies have manipulated both task and procedural variables and have, in many respects, been the most influential of all in improving our scientific understanding of tracking behavior because they have attempted, in a detailed and analytical fashion, to clarify the various response facets of tracking behavior and the variables determining them. It is perhaps safe to say that these studies have stood as a numerical minority in tracking research, and this is unfortunate because such information stands as the foundation of any systematic empirical and theoretical organizations of tracking behavior. Neither the studies of tracking qua tracking which have arisen out of the applied interests of engineering psychology, nor the studies of theoretical psychology where tracking tasks have been used as a matter of convenience, can progress very far until their findings are related to the complex characteristics of tracking behavior. Analytical tracking studies in this vein will be discussed in some detail in later sections of this paper.

Areas of Neglect

With some exceptions, engineering psychology and general experimental psychology have tended to gloss over three fundamental topics which must be given more attention if we are eventually to have the beginning of a theory of tracking behavior:

1. Tracking tasks have never been defined other than by convention. Early interests in tracking behavior arose out of applied situations where a continuously generated error quantity had to be nulled by continuous operator movements. Laboratory studies of tracking follow this applied tradition of a continuous task, although on occasion discrete displacements of the input signal have been used (Craig, 1949; Ellson, Hill, & Craig, 1949; Rund, Birmingham, Tipton, & Garvey, 1957; Searle & Taylor, 1948; Taylor & Birmingham, 1948; Vince, 1948b, 1949). An attempt must be made, at least in a preliminary way, to define the allowable variations in input, both in type and functional form, as well as the characteristics of the control system used for responding.

2. Not enough attention has been given to the emphasis (largely British) on a more detailed description of behavior in tracking. Recognition

must be given to the presence and interaction of several overt and intervening response and stimulus classes, and how these factors act to determine the characteristics of the measured motor response.

3. Relatively little interest has been expressed in multidimensional tracking tasks having two or more stimulus sources in the same or different sense modalities, and corresponding dimensions in the control system for response to each source. Most tracking research has been performed on one-dimensional tasks. The implications of various ways of organizing multiple inputs and the control systems for response to them need more formalization and research.

This paper will, in turn, discuss issues, problems, and research associated with each of these three areas.

Definition of Tracking

A one-dimensional tracking task will be defined by the following conditions:

1. An externally driven input signal defines an index of desired performance and the operator actuates the control system to maintain alignment of the output signal of the control system with the input signal. The discrepancy between the two signals is the error and the operator responds to null the error. Two basic types of tracking tasks are differentiated by how this error quantity is represented: (*a*) Pursuit Tracking. The display has two indicants. One is actuated by the input signal and the other is linked to the output signal of the control system. The two indicants are presented directly to the operator and he responds to null the error difference between them. (*b*) Compensatory Tracking. The error to be nulled is not the difference between two directly observed indicants primarily linked to the input and output signal as in pursuit tracking. Instead, the error observed in pursuit tracking is abstracted and used to actuate a single indicant in relation to a fixed reference. The operator's task, just as in pursuit tracking, is to null this error. The principal difference between pursuit and compensatory tracking is that with the latter the operator never observes the uncontaminated action of the input or output signal directly —only the error difference between them.

2. The input signal is time-based and independent of the operator's response, i.e., the task is *paced*. A paced task is distinguished from a self-paced task where stimulus changes are a function of operator responding (Adams, 1954).

3. The control system has constraints that enforce certain transitional courses of action on the human operator. Instead of being able to move the control from a given position to any other position, the operator must move through defined intervening states of the control system. For example, consider a one-dimensional visual tracking task using a pivoted control lever with hypothetical control Positions A, B, C, and D. If the operator is at Position B at time t, he has a three-choice decision for moving the control at time $t+1$, each with a probability of being correct: he can repeat the response of time t and leave the control at Position B, or he can move the control to either Position A or C. At the two extreme limiting positions of the control, only two choices are involved: leave the control where it is, or move it to the position adjoining the limiting one. By this definition, any task where the operator has free transitional access to all of the control system states is prohibited from being a tracking task.

4. The states of the input signal have the same transitional constraints

as the control system. The input signal, in changing from time t to $t+1$, must change according to constraints defined by the control system. By imposing the same constraints on the input signal and the control system, the tracking task is given a degree of feasibility for the human operator and means that the input cannot take any action which, in principle, cannot be met by action of the control system. This does not mean that a tracking task must allow near perfect performance by the operator. The input function may be a high frequency sine wave to which the operator can never achieve a high level of proficiency, but this is a behavioral matter and not a function of inherent design features of the task. Table 1 presents the permissible transitional states for the hypothetical four-state tracking task discussed above.

TABLE 1

Matrix Governing the Allowed Transitional States for the Input Signal and the Control System

		j			
		A	B	C	D
	A	Yes	Yes	No	No
	B	Yes	Yes	Yes	No
i	C	No	Yes	Yes	Yes
	D	No	No	Yes	Yes

Note—The matrix represents a hypothetical four-state one-dimensional tracking task. Cells marked with "Yes" indicate permissible transitions from the *i*th state at time t to the *j*th state at time $t+1$. "No" entries are absolute constraints and signify the denial of transition to a *j*th state from a prior *i*th state.

This definition is general and does not specify the characteristics of the input signal or the control system, other than indicating certain transitional restraints for both. The input states and the responses to them can be discrete or continuous, and the input can have any degree of regularity from nearly random (true randomness is denied by conditional restraints of the type shown in Table 1) to completely repetitive. The use of discrete states of the input signal deserves more than the passing attention it has been given in the past because they are particularly amenable to statistical structuring in terms of first and higher order probabilities (with the restraints noted). Another advantage of discrete inputs is that their duration is easily manipulable, making the number of events per unit of time an important dimension for investigation. This time variable has been termed the "speed or pacing factor" (Adams, 1954; Conrad, 1951, 1954; Wagner, Fitts, & Noble, 1954) and is analogous to number of cycles per second when a continuous input is used. One promising measure expressing the statistical coherency of a discrete input signal and the duration of its events is the informational measure of bits per unit of time (Shannon & Weaver, 1949). The rate of change, as well as higher derivatives, can also be a variable for discrete input events but no attempts have ever been made to explore these more complex dimensions.

The Complexity of Behavior in One-Dimensional Visual Tracking

The purpose of this section is to discuss some of the characteristics of the response classes which can be identified in one-dimensional visual tracking, as well as the issues surrounding them. Visual tracking will be analyzed because almost all tracking research has used the visual modality. However, in whatever broad empirical and theoretical conceptualizations of tracking behavior that might eventually mature, it will be necessary to structure the characteristics of tracking in other sense modalities to. But since other modalities such as audition have received only exploratory attention (Humphrey & Thompson, 1952a, 1952b, 1953), it seems unduly speculative at this time to include them.

Rather than the servotheory approach which has been the frame of reference of some investigators, an attempt will be made to demonstrate, on the basis of the available experimental evidence, that tracking behavior involves a linked chain of overt and internal stimuli and responses and is much more complex than implied by the prominent error-nulling characteristics of the servoanalogy. While the servoanalogy is adequate enough for its schematic purposes, the behavioral phenomena cannot be viewed so simply. There are three major areas for discussion: the observing response which orients receptors to sense stimulus events on the display, the prediction responses where the operator learns to anticipate future characteristics of the input signal, and the hypothesis that the measured motor response, even in continuous tracking, is intermittent and not smooth graded movements that might appear to a casual observer.

Most of the phenomena will be discussed in greatest detail under the heading of pursuit tracking, and the presence of the same or similar phenomena in compensatory tracking will, in most cases, be obvious. Behavioral considerations which are uniquely characteristic of compensatory tracking will be treated separately.

Pursuit Tracking

Observing Response

The sensing of the displayed indicants driven by the input and output signals, as well as the error difference between them, is by the observing response. These three environmental quantities each play an important role in pursuit tracking and their moment-to-moment state is sampled as the observing response orients the receptors to them. The input indicant is the desired state, the error difference between the input and output signal represents how well the desired state is achieved, and the output indicant gives knowledge of results on how specific sequences of motor movements are represented on the display. Some general attention has been given to the general role of the observing response (Wyckoff, 1952), but within the context of tracking it is considered as having two functions: head and/or eye movements to direct the visual receptors to spatially separated stimuli, and the discrimination of stimulus change. The head and/or eye movements can be considered overt aspects of the observing response and potentially measurable (Mackworth & Mackworth, 1958). However, the discrimination function of the observing response is an inferred phenomena, with its locus unspecified.

Common experience dictates the necessity for an observing response but there is also experimental evidence which documents its importance. Adams (1955), using the Rotary Pursuit Test, found that operations of repeatedly activating the visual observing response independently of the arm-hand goal response, and which presumably served to fatigue the observing response, resulted in a goal response decrement and permitted the inference that the performance level of the goal response is partly determined by the strength of the intervening observing response. Another relevant line of evidence is a study by Poulton (1952b) where it was found that pursuit tracking performance deteriorated when the two pointers on the display were increased in their spatial separation. One interpretation of this finding is that the greater spatial separation required more extensive orienting of the observing response with the result that less time, on the average, was devoted to each pointer. Viewing the observing response as the mechanism by which stimuli are sampled, the wider the spatial separation the less frequently each source of environmental stimuli is sampled and the less likely that an appropriate response will be made. Bearing on this sampling function of the observing response is a vigilance experiment by Jerison and Wallis (1957) where it was found that the scanning of three stimulus sources resulted in a lower rate of detecting aperiodic stimulus change than when only one source had to be watched.

Prediction Responses

The input signal in pursuit tracking actuates an indicant which is directly observed by the operator. To the extent that the operator can predict the regularities inherent in this input signal he will be able to anticipate the correct response movement and initiate at a time to minimize error. In the absence of a predictive capability the operator must wait for the change in the input signal to actually occur on the display, with the result that his response will generate tracking error as a function of a delay of at least one reaction time interval.

Helson (1949) and his associates, in their Foxboro studies of tracking during World War II, were perhaps the first to suggest that prediction behavior is manifest in reaction time values far less than those obtained in classical reaction time experiments. Bartlett (1951) has written an excellent paper on the role of anticipatory behavior which seems to be little known and referenced in the United States. The most extensive research on the prediction of directional course changes in the input signal has been by Poulton (1952a, 1952b, 1957a, 1957b, 1957c), and he distinguishes between two general classes of prediction: (*a*) receptor anticipation, which is analogous to the foreperiod of the classical simple reaction time experiment where a preparatory signal is presented to the operator in advance and establishes a "set" for response, and (*b*) perceptual anticipation, where no advance information is intentionally given each time but the operator nevertheless is able to predict the course of future signals on the basis of his past experience. It is this latter type of anticipation which is of greatest interest in tracking in that any knowledge of a future state of the input signal must be an acquired or learned prediction; the definition of a tracking task does not provide for foreknowledge of a state of the input signal. In one study (1952b) Poulton evaluated anticipation in pursuit tracking as a function of practice and two levels of input complexity—a simple harmonic motion and a complex harmonic course. Taking an anticipation of change in the input signal as a response of duration less than the expected reaction time of about .20 seconds, Poulton found that the subjects were predicting the simple harmonic course both early and late in practice, and that the success of prediction was a positive function of practice. Although overall tracking error decreased with practice on the complex input course, there was no evidence for improvement in anticipation and Poulton concluded that the improvement was largely attributable to increased manual dexterity. In this study, Poulton also investigated the smoothness of tracking, defined by the number of unnecessary discrete changes of speed that were made. The fewer the number of such changes, the better the performance. With the simple harmonic course, it was found that smoothness of response increased with practice but no such changes were found for the complex input. Poulton viewed his measure of smoothness as an additional index of anticipation because, when the operator was not anticipating, he would tend to wander off course and his tracking record would show a greater number of corrective movements. He observes that smoothness is a less sensitive measure of beneficial anticipation than response time because the operator may be tracking with a large lag but nevertheless tracking smoothly. Yet, the fact that the subjects tracked most smoothly for the harmonic input course which also produced the greatest degree of anticipation suggested that the covariation of these two measures reflects the same underlying ability to predict stimulus change in direction.

Another study by Poulton (1952a) used the same pursuit tracking apparatus as in his previous study (1952b) and investigated the accuracy with which an operator could predict the position of the input indicant for various amounts of time in the future. At the sound of a hammer blow the operator had to move the output indicant to the position anticipated for the input indicant when a bell sounded .50, 1.5, or 3.5 seconds later. This procedure was regularly repeated and resulted in a series of discrete responses predicting the position of the input indicant. The accuracy of prediction was better than chance for both simple harmonic and complex harmonic inputs, with the accuracy being greater for simple harmonic motion. On the basis of these experiments, Poulton concluded that course anticipation is an important determiner of the overall proficiency level in pursuit tracking. He hypothesized that higher input speeds place a greater premium on prediction because, as the speed of the input signal increases, the failure to anticipate means that a greater segment of the input course span will pass during the subject's reaction time period if he waits for stimulus change to actually occur before responding and a larger error will develop. An excellent review of the role of prediction in tracking and other types of visual-motor tasks has been published by Poulton (1957b).

A series of investigations by Gottsdanker (1952a, 1952b, 1955, 1956) is closely related to those of Poulton. Gottsdanker's studies were concerned with the prediction of velocities and accelerations of input rather than directional changes in the course, and were subsumed under the label prediction motion. The experimental approach required the subject to track a continuous input viewed through a narrow slit. The input was printed on paper in the form of parallel lines 5 millimeters apart, and the subject responded by trying to keep a pencil point between the two lines. He was told that when the input disappeared he was to project its path into the future as if he were attempting to follow an airplane that had gone behind a cloud. Some of the input paths had constant velocities but others had motions that were positively or negatively accelerated. In general, his findings show that constant velocities are accurately predicted, but that the prediction of accelerations tended to be of a constant velocity rather than the required increase or decrease in velocity. Gottsdanker interpreted this to mean that the subject responds on the basis of averages or integrations of preceding velocities. Two studies by Vince (1953, 1955) used a technique very similar to Gottsdanker's in investigations of what she termed "intellectual processes" in skilled performance. Another paper of interest on this topic, but not directly related to tracking, is by Leonard (1953).

The studies by Poulton and by Gottsdanker have involved the learning of prediction during the course of actual practice on a tracking task. A related line of investigation, which has been given some attention in another study by Poulton (1957a), is the effect of training to predict the stimulus source *prior* to actual motor practice in the total tracking task. This can be viewed as a part-whole transfer of training approach, where prediction responses are considered part of the response totality in tracking. Granting this, prediction responses should be trainable prior to whole-task practice and, in being a part of the total response complex, should have their strength reflected in the dependent motor response whose proficiency reflects the strength of all the response classes in the complex. This approach is quite similar to verbal pretraining methods where the operator is required to learn verbal responses to task stimuli prior to motor responses in the whole task (Arnoult, 1957; Goss, 1955). Although verbal pretraining studies have not dealt specifically with the problem of prediction responses, they are concerned with learned mediating responses where response produced stimuli are hypothesized to provide additional discriminative cues for the motor response (Goss, 1955; Osgood, 1953). Conceptually therefore, they appear quite similar to prediction responses and one might hypothesize that an adaptation of these same methods can be used for prior training in the prediction of input events in a tracking task. However, with our impoverished knowledge of the underlying nature of anticipatory mechanisms, it is plausible that prediction has nothing to do with mediating responses but, indeed, may be fundamentally a proprioceptive-oriented phenomenon. Giving proprioception a role in anticipatory behavior needs only the reasonable assumption that motor movements are conditioned to traces of proprioceptive stimuli and that, with practice, the occurrence of a proper configuration of proprioceptive stimuli will tend to elicit the next correct motor sequence. Certainly this is not to deny intellective processes or mediating responses as variables in prediction, but it does suggest that there might be at least two facets that deserve experimental inquiry. "Prediction response" is a commonly used label for anticipation in this paper but eventually it may prove to be a poor term if proprioception proves to be a paramount influence. The verbal pretraining studies throw the balance of the explanatory weight at present in the direction of mediating responses as the basis for anticipation, but definitive research on this topic remains to be done.

Characteristics of the Measured Motor Response

The basic nature of the motor movement activating the control system in a tracking task has been the subject of extensive discussion and controversy. The issue is whether the motor response is a continuous function of time or whether it is discontinuous and intermittent. The intermittency hypothesis stems from arguments that a responding effector has a period of refractoriness or reduced excitability before it can be made to respond in full strength again. Because this evidence stems from molar behavior data, it is called psychological refractory phase to distinguish it from the physiological refractory phase of individual nerve fibers. The similarity of psychological refractory phase and physiological refractory phase is in terms of reduced responsiveness following stimulation and response, but the levels of analysis of the two classes of phenomena are so different that it is perhaps safest to view them as analogous rather than stemming from a common underlying process.

Probably the first statement of psychological refractory phase was by Telford (1931) who found that reaction time to the second of a pair of auditory stimuli was lengthened if the time spacing of the two stimuli was reduced to .50 seconds, and he concluded that the subject becomes refractory in a manner comparable to the refractoriness of isolated nerves. Using Telford's study as a point of departure, Vince (1948a, 1948b) asked whether refractoriness is present in continuous tracking to give the motor response an intermittent, impulsive quality. She concluded that intermittent corrections every .50 seconds is a basic feature of human tracking responses in a manner quite comparable to Telford's finding for discrete stimuli and a reaction time response. If her interpretation is correct, the notion of psychological refractory phase becomes an important general principle. But in criticism of Vince's findings, psychological refractory phase refers to the periodicity of *motor movements* and not tracking *error*. Her conclusions were based on tracking error records and periodicities in them are not a function of motor movements alone but of the *difference* between the output signal generated by the motor movements and the input signal. Periodicities in the error function may be correlated with periodicities in motor movements but they are contaminated by the influence of the input signal and are not an unequivocal index on which to base conclusions about psychological refractory phase as a mechanism for inducing intermittent motor corrections.

With the exception of the foregoing studies by Vince, research on motor intermittence has been with discrete tasks, although many of the investigators have freely implied the generality of the phenomenon to include continuous tracking. Mainly, these studies conclude that reaction time to a second stimulus of a pair will be lengthened if the interstimulus interval is less than .50 seconds. A limit to this generalization is that very brief interstimulus intervals cause the stimuli to be perceived as a single entity, with the result that only a single response occurs. Vince (1948b, 1949) and Hick (1948) have both used discrete tracking tasks and have provided additional corroborative evidence on psychological refractory phase. Craik (1947, 1948) and Welford (1952) use these data for theoretical discussions on the generality of psychological refractory phase as a determinant of intermittency in responding. Poulton (1950) criticized the tendency to regard the refractory interval of .50 seconds as a human constant because the quasirandom presentation of stimuli did not allow the operator to form a proper preparatory set. When allowance is made for the acquisition of a preparatory set by having predictable stimuli, Poulton found that the refractory phase interval reduces to .20–.40 seconds. Davis (1956) and Elithorn and Lawrence (1955) also discuss the role of anticipatory set and the psychological refractory period. A general discussion of research and views on this topic is presented by Fitts (1951).

Another aspect to the intermittency hypothesis is that the duration of patterned movements to discrete stimuli can be less than visual-motor reaction time. This has implied response discontinuity to some investigators because the subject is executing response sequences momentarily independent of the magnitude of the visually perceived error. Because the response is not continuously guided by the primary visual tracking error quantity, it is, for a time, open-loop or intermittent (Searle & Taylor, 1948; Taylor & Birmingham, 1948). These authors conclude that response movements are under a kind of a "cam control" where the visually perceived error triggers a cammed sequence. On the basis of past experience the "cam" runs off the continuously varying force pattern, including starting and stopping, and all without visual or proprioceptive feedback.

Admitting the possibility that the continuous control of movements during an interval less than that required for visual-motor reaction time can be proprioceptive feedback, Chernikoff and Taylor (1952) conducted a study to see if kinesthetic reaction time was sufficient to account for the control of the response. They concluded that continuous tracking behavior is best described by the intermittency hypothesis, analogous to cam control where very brief movement sequences are run off in the absence of visual and proprioceptive guidance. Lashley (1951) in a parallel line of argument, is in agreement that kinesthetic reaction time cannot explain many facts of motor responding such as the finger movements of a skilled pianist moving at about 16 per second. These rates are too fast to allow kinesthetic feedback after each one, and Lashley postulates that some central sensory control is operating, presumably in a fashion similar to the cam hypothesis stated by Taylor and his associates. Craik (1947) holds a similar view. Arguing from piano playing to tracking is tenuous however, if for no other reason than that a musical composition provides foreknowledge of a requirement for movement sequences, and reaction time to each one is known to be greatly shortened under these special conditions (Vince, 1949). Advance notice of stimuli is not a characteristic of tracking tasks. Moreover, Poulton's work (e.g., 1952b) has shown that learning to anticipate stimulus sequences is revealed in greatly shortened reaction time values. It is hardly surprising that a trained musician can sometimes sidestep the restraints of an elementary afferent-efferent loop and receive guidance from learned, internal sources.

Gibbs (1954a, 1954b) in two important papers effectively argues against the hypothesis that continuous motor movements do not have continuous kinesthetic feedback guiding them. He points out that arguments based on kinesthetic reaction time fail to distinguish between the connecting and conducting functions of the central nervous system. Gibbs observes that kinesthetic reaction time to discrete stimuli can be considered the connecting time between kinesthetic stimulation and overt motor response, and this has little bearing on continuous kinesthetic or neural conduction during voluntary movement. Gibbs bases his discussion on physiological data by Matthews (1933) which showed that a muscle had "tension" afferents and "stretch" afferents which, respectively, provide sensing of static position and of movement of a limb.

Tension afferents respond primarily when the muscle is at rest and has an electrical discharge approximately proportional to the logarithm of the tension. Stretch afferents, on the other hand, respond when the muscle is stretched in movement and has a rate of electrical discharge proportional to the rate of stretch, and Gibbs holds that this is the source of *continuous* kinesthetic feedback monitoring. The subject must "know" limb position in guiding his movements and Gibbs holds that this is obtained by integrating the rate function. The notion of a finite integration period might suggest that Gibbs' hypothesis is essentially the same as the intermittency hypothesis because successive integrations might be revealed as intermittent movements of .50 seconds as limb position is successively "computed." Actually, the implications are quite different because Gibbs' hypothesis would seem to hold that there are conditions where an integration interval of .50 seconds would apply but that integration intervals of longer duration are equally possible. Gibbs' physiological hypothesis would seem to allow for perfectly smooth tracking movements of relatively long duration and, indeed, this is a common observation in tracking records. Oddly enough, relatively long periods of smooth responding in continuous tracking have not served as grounds for seriously challenging the intermittency hypothesis. Craik (1948) and Noble et al., (1955) remark on these smooth responses and offer the ad hoc explanation that intermittent movements are occurring in accordance with the principle of psychological refractory phase but that the subject's acquired capability to predict input sequences has overlaid a smoothing effect. While prediction responses may well have some sort of smoothing influence, it also may be true that the intermittence hypothesis is false for continuous tracking and that relatively long, smooth responses frequently occur in the absence of prediction behavior. Gibbs' work emphasizes the rather simple fact that the intermittency hypothesis has its validity derived from research on discrete tasks and its generalization to continuous tracking may be inappropriate.

Gibbs' use of Matthews' findings raises the interesting idea that proficiency in making accurate acclerations in tracking is related to the subject's ability to discriminate changes in the rate of kinesthetic impulses. One interpretation of Gottsdanker's findings (1952a, 1952b, 1955, 1956) that the subject poorly predicts velocity changes is that he cannot kinesthetically discriminate with enough accuracy those velocity changes which he visually perceives. However, this interpretation must be approached cautiously because it fails to consider that the inability to discriminate velocity changes conceivably could be on the visual-perceptual side rather than the kinesthetic. To interpret Gottsdanker's data properly we must, by independent operations, determine the relative capabilities of perceptual and kinesthetic discrimination of acceleration. If the operator cannot perceptually discriminate the velocity changes involved, then the motor response system is not receiving adequate information and the overt response cannot be expected to reflect information that has not been received. Or, conversely, the operator may be perfectly able to discriminate the velocity change perceptually but he may be unable to translate it into the proper accelerated movement because he cannot make sufficiently accurate kinesthetic discriminations.

Some work on the perceptual discrimination of instantaneous changes in velocity has been done by Hick (1950) and Brandalise and Gottsdanker (1959). Too, this general line of reasoning suggests the hypothesis that the relative effectiveness of position, rate, and acceleration tracking may be related to the compatibility of perceptual and kinesthetic events.

Compensatory Tracking

The discussion of research and problems under the heading of pursuit tracking applies also to compensatory tracking. Whatever differences exist are resident in the different ways in which the two types of tracking tasks have their data organized on the display. The presentation of only the error quantity in compensatory tracking means that performance usually will be poorer for two reasons:

1. The operator cannot see the input signal directly which means that he is handicapped in the acquisition of prediction responses.

2. The operator cannot see the output signal directly so he is handicapped in receiving knowledge of results. In addition to influencing the acquisition of simple visual-motor learning where prediction behavior is absent, this factor also influences the acquisition of prediction responses because the operator cannot unequivocally verify the results of any particular prediction response.

Depending upon task circumstances, some prediction behavior can be expected to form under compensatory tracking conditions. The error signal is a function of both the input and the output signals, and at times the regularities in the input will be discernible. Poulton (1952b) has shown that prediction behavior does occur with practice in compensatory tracking but that prediction is impressively superior in pursuit tracking. Undoubtedly this is one of the factors which almost always renders pursuit tracking superior to compensatory tracking (Hartman & Fitts, 1955; Poulton, 1952b).

Nor can we assume that the absence of a direct presentation of the output signal means that knowledge of results is completely absent. There is evidence from a study by Chernikoff and Taylor (1957) that when the input signal of a continuous tracking task is a low frequency input the subject receives fairly adequate knowledge of results, probably because the motor movements produce output frequencies which are higher than the input frequency changes. This is deduced from the slightly better performance that was found in this study for compensatory over pursuit tracking when the input was a low frequency signal. At higher frequencies, they found that pursuit tracking maintained its well-known advantage over compensatory tracking.

Two-Dimensional Tracking

A two-dimensional tracking task has two stimulus sources commanding response, with each source having its own separate input signal and a dimension of the control system for response. An example of a two-dimensional visual tracking task would be two voltmeter stimulus sources with a left-hand control lever for response to one and a right-hand lever for response to the other. Or, the two stimulus sources could have a bisensory distribution, with one visual and one auditory. Our ignorance of variables involved in the various ways to organize a two-dimensional tracking task dictates that only a limited examination of some of the issues be made. The discussion will be restricted to two cases: spatially separated visual sources, and bisensory sources where one is auditory and the other is visual. Nothing of importance is known of the effects of control system design as it bears on the distribution of the two response dimensions among the possible effector systems, so it will not be discussed. Nor will the relative advantages of pursuit and compensatory displays be discussed for whatever special implications might be found for two-dimensional tasks.

Because almost all of the research in tracking has employed one-dimensional visual tasks, it is unfortunately necessary to attempt this preliminary discussion of two-dimensional tasks on a rather thin foundation of empirical findings. Perhaps the dearth of analytical data on more complex tracking tasks is because of the implicit view of many psychologists that it is desirable to progress in research from simple to complex systems, and that the laws of complex systems will tend to fall into place once the relationships for simpler tasks are established. On the other hand, it is possible to defend the position that parallel law-seeking at two levels of analysis will result in two bodies of laws, each appropriate for its own domain. As these two bodies of knowledge develop, specific research can then be directed towards finding the empirical composition laws which express the interactions relating the laws of the two strata. If this view is allowed, it does not seem necessary that the study of multidimensional tracking tasks should await the codification of laws governing one-dimensional tracking.

To facilitate exposition, the following terminology has been adopted. Each stimulus source and its dimension of the control system will be called a component task of the total task. Response in the component task will be termed the component response of the total response. As before, the observing response will serve to orient the receptors to the events emitted by the stimulus sources.

Visual Tracking

Observing response. One of the distinguishing features of a two-dimensional visual tracking task is that there is not only the need for scanning *within* a source but also the more demanding requirement to scan *between* sources. This added response requirement is importantly a product of the task variable called load (Conrad, 1951, 1955). Load is defined as the number of stimulus sources and has an expected interaction with the rate of events emitted from each source. This latter variable has been termed speed (Conrad, 1951, 1954). Performance deteriorates both with increase in speed and load. Moreover, it has been shown that response proficiency is a function of the extent to which events in spatially distributed sources overlap in time and command two simultaneous responses (J. F. Mackworth & N. H. Mackworth, 1956; N. H. Mackworth & J. F. Mackworth, 1956, 1957). Another important task variable which would certainly interact with speed and load in determining the observing response is the amount of the spatial separation of the sources. Tracking proficiency as a function of the amount of spatial separation has not been systematically studied.

Prediction responses. An important but unverified implication for prediction responses in a two-dimensional visual task is that they might reduce the major requirement for visually scanning the stimulus sources and improve tracking performance. Prediction responses in a one-dimensional task are known to benefit motor performance. We might hypothesize that in a two-dimensional task there is not only prediction within each source but also prediction *between-sources*. If the human operator can learn to predict events within a source, it would seem that he might learn of the covariation between the events of the two sources. Given an event in one source he would have some likelihood of correctly predicting the concurrent event in the other source and consequently would not need to attend visually to this source as often. We know nothing of these matters but between-source prediction is a reasonable expectation.

Component response differentiation. Two-dimensional tracking often involves two or more component response effector systems, such as both hands or a hand and a foot, and this raises the issue of motor interaction between the two systems. It is a common observation that initial stages of total response in a multidimensional task are often typified by a level of uncoordinated activity and error far greater than might be expected from low habit strength in each component response separately. But, as practice proceeds, these interactions of component responses tend to drop out completely or show a marked decrease, with each participating component response effector system becoming smoothly proficient. This phenomenon shall be called component response differentiation. Within the framework of his S-R contiguity theory, Guthrie (1952) discusses the acquired differentiation of component responses:

> . . . reduction of habit to essentials makes many habits local responses no longer involving the whole body. When we are practiced we drive and talk, or play the piano and smoke, or skate and greet a friend at the same time. At first this is impossible because driving, playing, skating all include a mass of action that is not essential to the performance but is present because it is part of total associated complex bound together by conditioning. In time, many irrelevant movements are dropped out from the complex and the activity is limited to the muscles and the movements required for the performance. This process is, of course, never complete. Perfect grace, which means the use only of the essential muscles and this use only to the point necessary for the action, is only approximated, never reached (p. 109).

How component response interaction is manifested in a two-dimensional tracking task is not known at this time. However, the extensive literature on experimentally induced muscular tension, which has been organized by Meyer (1953) in terms of physiological hypotheses, leaves little doubt that interaction of simultaneous motor responses occurs. The concern of Meyer's review and analysis was the effects of experimentally induced muscular tension where usually a static, muscular tension-inducing component response accompanies a more central learning activity, such as rotary pursuit or paired-associates learning. The major area of interest for two-dimensional tracking, but where much less is known, concerns total tasks where all component tasks impose a learning requirement on their respective component responses. Perhaps, as Guthrie suggests, the interaction will all but disappear. But until a means of defining and measuring the course of component response differentiation in tracking is uncovered, there is no reason for discussion beyond this passing mention of a potentially important area.

Visual-Auditory Bisensory Tracking

The major issue for two-dimensional tracking with one visual and one auditory source is whether there is interaction which intrinsically pres vents the two stimulus event stream-from being processed simultaneously.

While we might intuitively surmise that a total task organized in this manner will be superior to two-dimensional visual tracking because each stimulus stream, with its own sense modality, gives the operator a higher load carrying capability in that he does not have to time-sample the sources with the observing response as he does in two-dimensional visual tracking, there is no evidence of the conditions for which this can be true, if at all. As a first experimental question it would seem desirable to attack the pure case in two-dimensional bisensory tracking and ask whether it is possible to *simultaneously* process two stimulus streams without impairing interaction effects. Any research program should have a strategy which sets up a hierarchy of research questions whose answers are ordered in terms of their contribution to the delineation of variables and laws, and in bisensory tracking the best strategy is suggested to be one of first determining whether the human operator can process two event streams at once. Having determined the empirical truth or falsity of this hypothesis, we will be in a better position to comparatively examine the relative merits of all-visual and bisensory tasks. Later variables to consider would be the differential capabilities of the visual and auditory senses for different classes of stimulus inputs (Henneman & Long, 1954).

Subjectively we all have the confident feeling that we can handle visual and auditory events simultaneously. It is commonplace to encounter the observation that one can simultaneously read a book and listen to the radio. Yet, as with most anecdotal accounts, they may be true but the absence of experimental controls precludes any proof of the thesis. Thus, an explanation of these experiences of everyday life is just as plausible in terms of rapid sensory shifting from one data stream to another. Experimentally, the issue is a delicate one and will require careful analysis and experimentation to decide it conclusively.

The experimental design necessary to prove or disprove that the human operator is a one-channel system must, as a minimum, show that performance of each component response in a bisensory tracking task will, after practice, be the same as performance when each component task is practiced out of total task context as a separate task. But what interpretation can be given if component response measures in bisensory tracking performance fail to achieve the level attained on part tasks? The hypothesis that the human operator is a single channel data processing system is supported but the investigator is then faced with the new question of the locus of the interaction. There are at least four possibilities which must then be resolved, although it will take some ingenuity and analysis to operationally differentiate them for laboratory testing:

1. The human operator is truly a one-channel system and, when two units of stimuli arise simultaneously, one must be temporarily stored while response occurs to the other. At the completion of the first response, the second stimulus unit is removed from central storage and response is made to it.

2. No storage is required. The operator is capable of simultaneously processing two event streams but there is motor interaction which prevents the two responses from simultaneously occurring with the same effectiveness that would be observed for any one of them separately. In effect, this hypothesis is consistent

with Guthrie's position that there is always some interaction between simultaneously functioning response systems, even after very large amounts of practice. Component response differentiation is never complete.

3. No storage is required and there is no interaction of responses at the motor level. However, there is sensory interaction which results in a degradation in performance that would be absent if only one stream of stimuli were being handled. Evidence for sensory interaction is presented in a number of papers (Child & Wendt, 1938; Gilbert, 1941; Gregg & Brogden, 1952; Hartman, 1934; London, 1954; Ryan, 1940).

4. Combinations of the above three possibilities.

The most relevant research on simultaneous bisensory data processing for tracking is by Davis (1957). While he did not study tracking or even strict simultaneity of bisensory events, he did study the effects of very small time intervals between a visual and an auditory stimulus and the experiment makes a significant contribution to the topic. Following the generalizations on psychological refractory phase, Davis asked whether the operator is refractory if the second of two successive stimuli impinges on a different sense modality than the first. Using the reaction time response and stimuli of very brief duration, Davis found that the reaction time to the second signal increased as the interstimulus interval decreased. The data show that the phenomenon which has come to be known as psychological refractory phase operates for two successive stimuli in two sense modalities about as it does for two successive events in a single sense modality. In some fashion, a "queuing of signals," to use the engaging phrase of Davis, occurs whether stimuli arrive over one or two sense channels. Davis finds his data consistent with a model of the human operator as a single channel information system. If we can assume that the processing of simultaneous events is a special zero-interval case of intervals for successive stimuli, the extrapolation of the Davis findings to the zero interstimulus interval suggests a substantial impairment in performance. An empirical study of truly simultaneous events must be done but the Davis experiment is unquestionably provocative on the simultaneity issue.

SUMMARY AND CONCLUSIONS

This paper has reviewed some of the major issues and problems in the study of human tracking behavior. Apart from the complexities that are inherent in the analysis of closed-loop behavior, which is somewhat more complicated than the open-loop situations used by most psychologists in their studies of human behavior, tracking behavior is beset with the added complications of mediating responses and stimuli which are important variables intervening between the display and the measured motor response. Moreover, all of these variables assume further complications when they are cast in the matrix of multidimensional tracking tasks with two or more stimulus sources, each with a corresponding dimension of the control system for response to them. And, not only do multidimensional tasks have complications resulting from a compounding of the effects of variables found in one-dimensional tracking, but they have the added issues of how one or more sense modalities process the incoming data and how the component response systems interact throughout learning to become partly or completely noninteractive (differentiated). We appear to be a long way from understanding these factors and, until we do, we are a long way from the beginnings of any kind of theory of tracking. British research has been most influential in illuminating the characteristics of tracking behavior, with its experimental examination of what is learned (e.g., prediction behavior in tracking), and its study of the intermittency hypothesis. This approach of British investigators would seem to be mandatory for our eventual theoretical description of tracking, and is in some contrast to the approach of engineering psychologists in the United States who tend to emphasize measures of tracking behavior as a function of task variables and often bypass detailed analyses of the learned behavior. Some important exceptions to this emphasis on the domestic scene has been the early work of the Naval Research Laboratory, Gottsdanker, and recent work by Briggs and his associates on learning and transfer as a function of task variables.

If this paper can be said to have a point of view, it is that tracking research is in need of a rapprochement of the interests of the engineering psychologist, with his focus on task variables and the measurement of time-based behavior, and interests of the traditional experimental psychologist who tends to emphasize behavior as a function of variables which determine conceptual states such as habit, work inhibition, motivation, mediating responses, etc. Physical servotheory has been a prominent attempt in engineering psychology to describe tracking behavior, but the absence of variables defining conceptual states long known to influence behavior eliminates it as a psychological theory of any stature, quite apart from its formal shortcomings for the description of nonlinear human behavior. It is unlikely that a theory of tracking behavior will emerge until these conceptual variables are included, along with time series measurement and task variables which traditionally have occupied engineering psychology.

REFERENCES

Adams, J. A. The effect of pacing on the learning of a psychomotor response. *J. exp. Psychol.*, 1954, **47**, 101–105.

Adams, J. A. A source of decrement in psychomotor performance. *J. exp. Psychol.*, 1955, **49**, 390–394.

Adams, J. A. Some implications of Hull's theory for human motor performance. *J. gen. Psychol.*, 1956, **55**, 189–198.

Adams, J. A. The relationship between certain measures of ability and the acquisition of a psychomotor criterion response. *J. gen. Psychol.*, 1957, **56**, 121–134.

Adams, J. A., & Reynolds, B. Effect of shift in distribution of practice conditions following interpolated rest. *J. exp. Psychol.*, 1954, **47**, 32–36.

Arnoult, M. D. Stimulus predifferentiation: Some generalizations and hypotheses. *Psychol. Bull.*, 1957, **54**, 339–350.

Bahrick, H. P. An analysis of stimulus variables influencing the proprioceptive control of movements. *Psychol. Rev.*, 1957, **64**, 324–328.

Bahrick, H. P., Bennett, W. F., & Fitts, P. M. Accuracy of positioning responses as a function of spring loading in a control. *J. exp. Psychol.*, 1955, **49**, 437–444.

Bahrick, H. P., Fitts, P. M., & Schneider, R. Reproduction of simple movements as a function of factors influencing proprioceptive feedback. *J. exp. Psychol.*, 1955, **49**, 445–454.

Bartlett, F. C. Anticipation in human performance. In G. Ekman, T. Husén, G. Johansson, & C. I. Sandström (Eds.), *Essays in psychology*. Uppsala: Almquist & Wiksells, 1951. Pp. 1–17.

Birmingham, H. P., & Taylor, F. V. A design philosophy for man-machine systems.

Proc. IRE, 1954, **42**, 1748–1758.

Bower, J. L., & Schultheiss, P. M. *Introduction to the design of servomechanisms.* New York: Wiley, 1958.

Brandalise, B. B., & Gottsdanker, R. M. The difference threshold of the magnitude of visual velocity. *J. exp. Psychol.*, 1959, **57**, 83–88.

Briggs, G. E., Bahrick, H. P., & Fitts, P. M. The effects of force and amplitude cues on learning and performance in a complex tracking task. *J. exp. Psychol.*, 1957, **54**, 262–268.

Briggs, G. E., & Fitts, P. M. Tracking proficiency as a function of visual noise in the feedback loop of a simulated radar fire control system. *USAF Personnel Train. Res. Cent. res. Rep.*, 1956, No. AFPTRC-TN-56-134.

Briggs, G. E., Fitts, P. M., & Bahrick, H. P. Learning and performance in a complex tracking task as a function of visual noise. *J. exp. Psychol.*, 1957, **53**, 379–387.

Briggs, G. E., Fitts, P. M., & Bahrick, H. P. Transfer effects from a single to a double integral tracking system. *J. exp. Psychol.*, 1958, **55**, 135–142.

Brown, G. S., & Campbell, D. P. *Principles of servomechanism.* New York: Wiley, 1948.

Chernikoff, R., & Taylor, F. V. Reaction time to kinesthetic stimulation resulting from sudden arm displacement. *J. exp. Psychol.*, 1952, **43**, 1–8.

Chernikoff, R., & Taylor, F. V. Effects of course frequency and aided time constant on pursuit and compensatory tracking. *J. exp. Psychol.*, 1957, **53**, 285–292.

Child, I. L., & Wendt, G. R. The temporal course of the influence of visual stimulation upon the auditory threshold. *J. exp. Psychol.*, 1938, **23**, 109–127.

Conklin, J. E. Effect of control lag on performance in a tracking task. *J. exp. Psychol.*, 1957, **53**, 261–268.

Conrad, R. Speed and load stress in a sensori-motor skill. *Brit. J. industr. Med.*, 1951, **8**, 1–7.

Conrad, R. Speed stress. In W. F. Floyd & A. T. Welford (Eds.), *Human factors in equipment design.* London: Lewis, 1954. Pp. 95–102.

Conrad, R. Some effects on performance of changes in perceptual load. *J. exp. Psychol.*, 1955, **49**, 313–322.

Craig, D. R. Effect of amplitude range on duration of responses to step function displacements. *USAF Air Materiel Command tech. Rep.*, 1949, No. 5913.

Craik, K. J. W. Theory of the human operator in control systems: I. The operator as an engineering system. *Brit. J. Psychol.*, 1947, **38**, 56–61.

Craik, K. J. W. Theory of the human operator in control systems: II. Man as an element in a control system. *Brit. J. Psychol.*, 1948, **38**, 142–148.

Davis, R. The limits of the "psychological refractory period." *Quart. J. exp. Psychol.*, 1956, **8**, 24–38.

Davis, R. The human operator as a single channel information system. *Quart. J. exp. Psychol.*, 1957, **9**, 119–129.

Elithorn, A., & Lawrence, C. Central inhibition: Some refractory observations. *Quart. J. exp. Psychol.*, 1955, **7**, 116–127.

Ellson, D. G. The application of operational analysis to human behavior. *Psychol. Rev.*, 1949, **56**, 9–17.

Ellson, D. G., Hill, H., & Craig, D. R. The interaction of responses to step function stimuli: II. Equal opposed steps of varying amplitude. *USAF Air Materiel Command tech. Rep.*, 1949, No. 5911.

Fitts, P. M. Engineering psychology and equipment design. In S. S. Stevens (Ed.), *Handbook of experimental psychology.* New York: Wiley, 1951. Pp. 1287–1340.

Fleischman, E. A. Dimensional analysis of psychomotor abilities. *J. exp. Psychol.*, 1954, **48**, 437–454.

Fleishman, E. A. A comparative study of aptitude patterns in unskilled and skilled psychomotor performance. *J. appl. Psychol.*, 1957, **41**, 263–272. (a)

Fleishman, E. A. Factor structure in relation to task difficulty in psychomotor performance. *Educ. psychol. Measmt.*, 1957, **17**, 522–532. (b)

Fleischman, E. A. Dimensional analysis of movement reactions. *J. exp. Psychol.*, 1958, **55**, 438–453.

Fleishman, E. A., & Hempel, W. E. Changes in factor structure of a complex psychomotor test as a function of practice. *Psychometrika*, 1954, **19**, 239–252.

Fleishman, E. A., & Hempel, W. E. The relation between abilities and improvement with practice in a visual discrimination reaction task. *J. exp. Psychol.*, 1955, **49**, 301–312.

Fleishman, E. A., & Hempel, W. E. Factorial analysis of complex psychomotor performance and related skills. *J. appl. Psychol.*, 1956, **40**, 96–104.

Floyd, W. F., & Welford, A. T. *Symposium on fatigue.* London: Lewis, 1953.

Forbes, T. W. Auditory signals for instrument flying. *J. aero. Sci.*, 1946, **13**, 255–258.

Gibbs, C. B. The continuous regulation of skilled response by kinaesthetic feedback. *Brit. J. Psychol.*, 1954, **45**, 24–39. (a)

Gibbs, C. B. Movement and force in sensorimotor skill. In W. F. Floyd & A. T. Welford (Eds.), *Human factors in equipment design*. London: Lewis, 1954. Pp. 103–117. (b)

Gilbert, G. M. Inter-sensory facilitation and inhibition. *J. gen. Psychol.*, 1941, **24**, 381–407.

Goode, H. H., & Machol, R. E. *System engineering*. New York: McGraw-Hill, 1957.

Goss, A. E. A stimulus-response analysis of the interaction of cue-producing and instrumental responses. *Psychol. Rev.*, 1955, **62**, 20–31.

Gottsdanker, R. M. The accuracy of prediction motion. *J. exp. Psychol.*, 1952, **43**, 26–36. (a)

Gottsdanker, R. M. Prediction-motion with and without vision. *Amer. J. Psychol.*, 1952, **65**, 533–543. (b)

Gottsdanker, R. M. A further study of prediction-motion. *Amer. J. Psychol.*, 1955, **68**, 432–437.

Gottsdanker, R. M. The ability of human operators to detect acceleration of target motion. *Psychol. Bull.*, 1956, **53**, 477–487.

Gregg, L. W., & Brogden, W. J. The effect of simultaneous visual stimulation on absolute auditory sensitivity. *J. exp. Psychol.*, 1952, **43**, 179–186.

Guthrie, E. R. *The psychology of learning*. (Rev. ed.) New York: Harper, 1952.

Hartman, B. O. The effect of target frequency on compensatory tracking. *USA Med. Res. Lab. Rep.*, 1957, No. 272.

Hartman, B. O., & Fitts, P. M. Relation of stimulus and response amplitude to tracking performance. *J. exp. Psychol.*, 1955, **49**, 82–92.

Hartmann, G. W. The facilitating effect of strong illumination upon the discrimination of pitch and intensity differences. *J. exp. Psychol.*, 1934, **17**, 813–822.

Helson, H. Design of equipment and optimal human operation. *Amer. J. Psychol.*, 1949, **62**, 473–497.

Henneman, R. H., & Long, E. R. A comparison of the visual and auditory senses as channels for data presentation. *USAF WADC tech. Rep.*, 1954, No. 54-363.

Hick, W. E. Discontinuous functioning of the human operator in pursuit tasks. *Quart. J. exp. Psychol.*, 1948, **1**, 36–57.

Hick, W. E. The threshold for sudden changes in the velocity of a seen object. *Quart. J. exp. Psychol.*, 1950, **2**, 33–41.

Howland, D., & Noble, M. E. The effect of physical constants of a control on tracking performance. *J. exp. Psychol.*, 1953, **46**, 353–360.

Hull, C. L. *Principles of behavior*. New York: Appleton-Century, 1943.

Humphrey, C. E., & Thompson, J. E. Auditory Display: II. Comparison of auditory and visual tracking in one dimension: A. Discontinuous signals, simple course. *Johns Hopkins U. Appl. Physics Lab. Rep.*, 1952, No. TC-146. (a)

Humphrey, C. E., & Thompson, J. E. Auditory Display: II. Comparison of auditory tracking with visual tracking in one dimension: B. Discontinuous signals, complex course. *Johns Hopkins U. Appl. Physics Lab. Rep.*, 1952, No. TG-147. (b)

Humphrey, C. E., & Thompson, J. E. Auditory Display: II. Comparison of auditory tracking with visual tracking in one dimension: C. Continuous signals, simple intermediate and complex courses. *Johns Hopkins U. Appl. Physics Lab. Rep.*, 1953, No. TG-194.

Jerison, H. J., & Wallis, R. A. Experiments on vigilance one-clock and three-clock monitoring. *USAF WADC tech. Rep.*, 1957, No. 57–206.

Kimble, G. A., & Horenstein, Betty R. Reminiscence in motor learning as a function of length of interpolated rest. *J. exp. Psychol.*, 1948, **38**, 239–244.

Lashley, K. S. The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior*. New York: Wiley, 1951.

Leonard, J. A. Advance information in sensori-motor skills. *Quart. J. exp. Psychol.*, 1953, **5**, 141–149.

Lewis, D. Motor skills learning. Symposium on psychology of learning basic to military training problems. *Res. Develpm. Bd., Comm. Hum. Resour., Rep.*, 1953, No. HR-HTD-201/1.

London, I. D. Research on sensory interaction in the Soviet Union. *Psychol. Bull.*, 1954, **51**, 531–568.

Mackworth, Jane F., & Mackworth, N. H. The overlapping of signals for decisions. *Amer. J. Psychol.*, 1956, **69**, 26–47.

Mackworth, Jane F., & Mackworth, N. H. Eye fixations recorded on changing visual scenes by the television eye-marker. *J. Opt. Soc. Amer.*, 1958, **48**, 439–445.

Mackworth, N. H., & Mackworth, Jane F. Visual search for successive decisions. *Med. Res. Council*, 1956, No. APU 234/56.

Mackworth, N. H., & Mackworth, Jane F. Temporal irregularity in a multi-source task. *Med. Res. Council*, 1957, No. APU 264/57.

MATTHEWS, B. H. C. Nerve endings in mammalian muscle. *J. Physiol.*, 1933, **78**, 1–53.

MELTON, A. W. (Ed.). *Apparatus tests.* Washington: United States Government Printing Office, 1947. (*AAF Aviat. Psychol. Prog. res. Rep.* No. 4)

MEYER, D. R. On the interaction of simultaneous responses. *Psychol. Bull.*, 1953, **50**, 204–220.

NOBLE, M. E., FITTS, P. M., & WARREN, C. E. The frequency response of skilled subjects in a pursuit tracking task. *J. exp. Psychol.*, 1955, **49**, 249–256.

OSGOOD, C. E. *Method and theory in experimental psychology.* New York: Oxford Univer. Press, 1953.

PAYNE, R. B., & HAUTY, G. T. The effects of experimentally induced attitudes upon task proficiency. *J. exp. Psychol.*, 1954, **47**, 267–273.

POULTON, E. C. Perceptual anticipation and reaction time. *Quart. J. exp. Psychol.*, 1950, **2**, 99–112.

POULTON, E. C. The basis of perceptual anticipation in tracking. *Brit. J. Psychol.*, 1952, **43**, 295–302. (a)

POULTON, E. C. Perceptual anticipation in tracking with two-pointer and one-pointer displays. *Brit. J. Psychol.*, 1952, **43**, 222–229. (b)

POULTON, E. C. Learning the statistical properties of the input in pursuit tracking. *J. exp. Psychol.*, 1957, **54**, 28–32. (a)

POULTON, E. C. On prediction in skilled movements. *Psychol. Bull.*, 1957, **54**, 467–478. (b)

POULTON, E. C. On the stimulus and response in pursuit tracking. *J. exp. Psychol.*, 1957, **53**, 189–194. (c)

RUND, P. A., BIRMINGHAM, H. P., TIPTON, C. L., & GARVEY, W. D. The utility of quickening techniques in improving tracking performance with a binary display. *USN Res. Lab. Rep.*, 1957, No. 5013.

RYAN, T. A. Interrelations of the sensory systems in perception. *Psychol. Bull.*, 1940, **37**, 659–698.

SEARLE, L. V., & TAYLOR, F. V. Studies of tracking behavior: I. Rate and time characteristics of simple corrective movements. *J. exp. Psychol.*, 1948, **38**, 615–631.

SHANNON, C. E., & WEAVER, W. *The mathematical theory of communication.* Urbana: Univer. Illinois Press, 1949.

SIDDALL, G. J., & ANDERSON, D. M. Fatigue during prolonged performance on a simple compensatory tracking task. *Quart. J. exp. Psychol.*, 1955, **7**, 159–165.

TAYLOR, F. V. Psychology and the design of machines. *Amer. Psychologist.*, 1957, **12**, 249–258.

TAYLOR, F. V., & BRIMINGHAM, H. P. Studies of tracking behavior: II. The acceleration pattern of quick manual corrective responses. *J. exp. Psychol.*, 1948, **38**, 783–795.

TAYLOR, F. V., & GARVEY, W. D. The limitations of a 'Procrustean' approach to the optimization of man-machine systems. *Ergonomics*, 1959, **2**, 187–194.

TELFORD, C. W. Refractory phase of voluntary and associative responses. *J. exp. Psychol.*, 1931, **14**, 1–35.

VINCE, MARGARET A. Corrective movements in a pursuit task. *Quart. J. exp. Psychol.*, 1948, **1**, 85–103. (a)

VINCE, MARGARET A. The intermittency of control movements and the psychological refractory period. *Brit. J. Psychol.*, 1948, **38**, 149–157. (b)

VINCE, MARGARET A. Rapid response sequences and the psychological refractory period. *Brit. J. Psychol.*, 1949, **40**, 23–40.

VINCE, MARGARET A. The part played by intellectual processes in a sensori-motor performance. *Quart. J. exp. Psychol.*, 1953, **5**, 75–86.

VINCE, MARGARET A. The relation between hand movements and intellectual activity in a skilled task. *Quart. J. exp. Psychol.*, 1955, **7**, 82–90.

WAGNER, R. C., FITTS, P. M., & NOBLE, M. E. Preliminary investigations of speed and load as dimensions of psychomotor tasks. *USAF Personnel Train. Res. Cent. tech. Rep.*, 1954, No. AFPTRC-TR-54-45.

WARRICK, M. J. Effect of transmission-type control lags on tracking accuracy. *USAF tech. Rep.*, 1949, No. 5916.

WEISS, B. The role of proprioceptive feedback in positioning responses. *J. exp. Psychol.*, 1954, **47**, 215–224.

WELFORD, A. T. The "psychological refractory period" and the timing of high-speed performance: A review and a theory. *Brit. J. Psychol.*, 1952, **43**, 2–19.

WYCKOFF, L. B. The role of observing responses in discrimination learning. Part I. *Psychol. Rev.*, 1952, **59**, 431–442.

(Received October 2, 1959)

Psychological Bulletin
1961, Vol. 58, No. 1, 80-86

THE MATCHING PROBLEM WITH MULTIPLE JUDGES AND OBJECTS

DONALD W. FISKE[1]
University of Chicago

The purposes of this paper are to consider some difficulties involved in matching problems with multiple judges and objects, and to present some appropriate techniques for the analysis of such data. The matching problem is the problem of evaluating the accuracy of a set of judgments about a series of objects. In the usual form of the problem, a judge places each object into one of several specified and unordered categories. Since the number of categories is finite and is ordinarily small, each such set of judgments has an appreciable probability of occurrence by chance. It will be obvious that there is always an external or a priori criterion for scoring each judgment as correct or incorrect. As Mosteller and Bush indicate (1954, pp. 307–308), the matching problem is present in many apparently different designs which call for identifying, diagnosing, or otherwise classifying objects, persons, or responses. Examples are guessing the order of a deck of ESP cards, diagnosing clinical cases, and identifying the products of each of several designated persons.

The several questions asked by the experimenter may include: Can these objects be classified by these judges with better-than-chance success? If so, are some judges more successful than others, and are some objects more successfully classified than other objects? It will be seen that these questions are not specific to the matching problem: e.g., they may also be asked in the analysis of responses to an aptitude test. What is distinctive about the matching problem is that the number of categories is usually larger than two, that each is of equal interest (as contrasted to the usual psychometric preoccupation with "right" answers), and that the number of objects classified by each judge is usually small. When the judge makes judgments about many objects in each of the true categories, psychometric scoring methods provide scores that can be analyzed by familiar statistical techniques.

The matching problem takes several forms, depending primarily on the number of categories into which the objects fall. The number may be from two to O (the number of objects). Another variable in such designs is the judge's information concerning the distribution of cases over categories: e.g., he may have some prior knowledge which would constrain his judgments, or he may be told how many objects fall in each category; if the categories are male and female, he would put approximately 50% of the objects in each category, or he might be informed that exactly 50% were of each sex. For brevity, this paper will be limited to the general form of the problem in which the instructions do not fix the distribution of judgments and in which the objects are from a specified class. Thus the task may be to judge whether or not each object has a certain property, or the analysis may be limited to evaluating the accuracy of judgments about objects in a given

[1] I am indebted to many colleagues for their criticisms and suggestions on early drafts of this paper, especially: Desmond S. Cartwright, Lee J. Cronbach, Lyle V. Jones, Jack Sawyer, Charles van Buskirk, and Virgil Willis.

category, these having been presented along with objects from other categories.

The One-Variable Case

Consider the limited case in which one judge, *j*, makes judgments about each of a set of *O* randomly selected objects. The experimental question is whether his success is significantly greater or less than that which would be expected from chance matching (due to his ability, his biases, systematic errors, etc.). Statistical techniques have been developed and presented by various writers. Mosteller and Bush (1954) give the most comprehensive account. Also pertinent are papers by Chapman (1934, 1935, 1936), Dudek (1952), McHugh and Apostolakos (1959), Roberts (1958), and Vernon (1936a, 1936b, 1936c). Mathematical treatments and further references are given by Battin (1942), Cochran (1950), Gilbert (1956), Stevens (1938), and Wilks (1943, pp. 208–213).

The conclusions from this design obviously apply only to Judge *j* and the population from which the *O* objects were drawn. Such a study would be of value if positive results permitted the conclusion that there is at least one judge who can correctly judge objects of this type. Negative results would have no value unless there were something distinctive about the one judge.

In parallel fashion, one object, *o*, might be judged by a set of *J* judges, with the conclusions applying to that one object and the population from which the judges were drawn. While such an experiment would be worthwhile if the object had special significance, inability to generalize to other objects would usually make the study of little value.

The necessity for representative designs in studies of this sort has been pointed out by various writers (Crow, 1954, 1957; Hammond, 1954; Secord, 1952). (For a pertinent discussion of the problems of interpreting the results of matching studies, see Cronbach, 1948.)

The Two-Variable Case

The one-variable case can be replcated with *J* different judges, each being assigned a random sample of objects, the several samples being obtained independently. If a *p* value is obtained for each judge, the findings for the several judges can be pooled (e.g., through the chi square transformation—see Jones & Fiske, 1953). If the objects are randomly chosen, one can make inferences about judgments of objects in the population from which they were drawn. If the judges are also randomly drawn, inferences can be made to the population from which they came. Otherwise, the inferences must be limited to the particular objects or the particular judges, respectively. This design is suitable for testing for nonrandomness of judgments, but does not permit a comparison between objects. It is not optimal for testing for differences between judges since differences between samples of objects would contribute to apparent differences between judges.

In another design, the *O* objects are randomly assigned to the *J* judges ($J=O$), each judge making one judgment and each object being judged once. This appears to be an excellent design for testing whether judges of a certain kind can judge correctly about objects of a specified type: a given amount of judge effort is spread over the largest possible number of objects so that the errors of sampling judges and objects would tend to be minimized. The resulting data can be analyzed by an appropriate statistical test from those in the references cited above: e.g., a test to determine whether the obtained proportion of hits departs significantly from that expected on the basis of chance.

Such a design is ordinarily not used because it does not permit analysis of judge differences, object differences, or judge-object interactions. It is also not economical of the experimenter's effort insofar as each judge must be instructed or trained and each object must be prepared for presentation to the judges.

In a more common design, each judge judges each object. The resulting data can be recorded in a bivariate table, the rows indicating objects, the columns the judges, and the entry indicating success or failure (e.g., 1 or 0).

But the several observations in each column may not be independent of each other, and neither are those in each row—just as, in the restricted case, the results apply only to the one judge or to the one object. Therefore, the set of *JO* observations cannot be treated as independent. To test the departure of the grand mean of the *JO* observations from chance expectancy, an error term is needed. The variance of the *JO* distribution is unsatisfactory because its actual degrees of freedom are not known. This dependence among observations is the crucial problem in most matching experiments.

This critical problem has been seriously slighted or overlooked in previous work. Mosteller and Bush (1954, p. 311) combine results for several judges on one set of objects without reference to the consequent restriction on the conclusion. Vernon (1936c) pointed out that significant differences between judges or between objects would introduce a marked bias in his method for analyzing matching data. He therefore proposed that, where such differences are known or suspected, the data for the average judge (or average object) be the basis for inference. This suggestion would seem to involve throwing out most of the available information and to increase the danger of making a Type II error. For example, suppose that each judge does better than chance but the data for the average judge does not reach the selected level of significance; the results for the several judges taken together might still attain significance.

When each judge judges each object only once, there is no satisfactory direct test of the observations as a totality. If the nature of the material permits each judge to make a number of judgments about each object, each judgment being truly independent of all previous judgments for that object, his score for that object can be stated as a proportion and an overall test could be developed on the basis of the discrepancies between these proportions and those predicted from the proportions for the given row and column. It is ordinarily not possible to obtain such a series of independent judgments.

However, indirect tests using the judge or object means can be made. The experimenter can test whether the overall proportions for the several objects (each proportion being the mean success per object) differ from chance. He can also make a test of the set of mean successes per judge. These will be considered below. It should be noted here that while these two tests are of the same grand mean, they may lead to different conclusions because the variance of the judge means is ordinarily different from that of the object means.

There is a special case of this multiple judge and multiple object design in which one assumption of random sampling is not made: examples are studies of particular judges who are distinguished by their obvious expertness, or studies of a finite and small class of objects of special interest. Here the experimental question concerns the performance of these judges or judgability of these objects.

Their role in the design is analogous to that of the fixed constants in one model for analysis of variance, or to the operations or instrument for measurement in research in general. The conclusions are limited to the instrument since no random sampling is involved in the selection.

In such designs, the appropriate procedure is to obtain a score for each application of the instrument, and to test the mean of these scores. For example, one might determine whether a group of experts can make a blind diagnosis of each of a random set of cases, the score being the number of experts who correctly diagnosed each case. The mean number right would be compared with the mean expected on a chance basis (as derived from the actual distribution of judgments over categories), the error term being based on the distribution of scores for the several cases.

An alternative technique is to determine a *p* value for each case, and then combine the set of such values for the whole set of cases. The expected proportion is the proportion of all judgments which fall in the category to which the case belongs. Given this value, the probability of obtaining or exceeding the observed number of successful judgments for the case can be found in tables for the binomial distribution. The *p* values for the several cases can be pooled by the chi square transformation of the several *p* values (Jones & Fiske, 1953). This method assumes that the degree of success of the judges on one case does not affect the degree of success on any other case.

A refined method for evaluating the performance of each judge has been suggested to the author by Lee J. Cronbach in a personal communication. The set of *O* judgments made by each judge can be scored in the usual way by counting the number of hits or correspondences with the criterion classifications or identifications. Then this same set of judgments by this *j*, treated as a distribution, is randomly assigned to the several *o*'s, and the number of hits noted. This process is repeated for a number of trials sufficiently large to provide a sampling distribution from which one can estimate the chance probability of obtaining the actual number of hits earned by this *j*. This method takes into account the judge's biases or preferences for certain categories.

For an oversimplified illustration, suppose that four *o*'s have the actual classification of A, B, B, C, and suppose *j* judges them to be A, C, B, A, respectively, giving him two hits. Since the number of *o*'s is small, one can in this case determine the exact probability of two or more hits by comparing with the criterion order (A, B, B, C) each of the 12 possible orders of two A's, one B, and one C. We find that 4 of the 12 yield two or more hits and hence the probability of this judge making two or more hits is .33. This probability is lower than the expected probability based on the assumption that the probability is $\frac{1}{3}$ of a hit on each *o*.

The same method could be applied to each object, with the resulting values of *p* for the *O* objects being evaluated as a set. This would ordinarily be less appropriate than the approach based on the judges, because it is known that individual judges have response sets and other biases which should be taken into account. The extent to which judgments about a particular *o* are biased will probably be of smaller magnitude and will typically be of lesser interest. The selection of approach should be based primarily on one's objective: if the objects are viewed as the instrument tested, *p* values are obtained for judges and an inference is made about the sampled population of judges.

This method has the same type of

rationale as that employed by Crow (1954) to evaluate judges' predictions of the responses of each of a set of patients. In this Random Comparison Method of Chance Determination, the distribution of predictions for all judges on all objects was compared to the actual behavior of each object and the resulting sets of discrepancy scores were pooled. It was then possible to determine for how many of the objects the given judge had done better than the median for all possible judge-object pairings. (For a study comparing correspondence between two records from each case with correspondence between paired records from different cases, see Kelly & Fiske, 1951, pp. 135–140.)

Up to this point, only designs for testing the obtained mean success have been considered. It should be noted that such means may be significantly below (as well as above) the value expected from chance matching due to systematic errors in judgment. A complete analysis of such data would also test whether the variance of the obtained scores was significantly different from that expected with chance matching, since significant differences between judges or between objects, or significant judge-object interaction may be present, even when the mean is not significantly different from chance.

It must be emphasized again that in all matching studies involving multiple judgments about a set of objects, the several observations cannot be treated as independent and the experimental design must take explicit cognizance of this restricting condition.

Testing for Differences between Judges or between Objects

The question of differences between the performances of judges or differences between objects in the ease with which they are correctly judged is often of interest, even when the nonrandomness of the judgments has not been tested. If nonrandomness has been tested, the question of individual differences may be of interest even when the mean does not depart significantly from chance.

Such testing for differences would be straightforward if the observations were a continuous variable: e.g., if each object could be judged at several different times by each judge. When the data are recorded as discrete observations (1 or 0), techniques such as conventional analysis of variance are not suitable.

One can, however, use the analysis of variance of ranks or *W*, the coefficient of concordance (Kendall, 1948). To test for differences between objects, we would rank the *O* values for each judge—i.e., we would assign the appropriate (tied) rank to all successes and similarly with the failures, making the appropriate adjustment for the ties. As in all designs where the objects are judged by all judges, it would be necessary to have the order of judgment randomized to control for effects of position on success.

Also appropriate is the chi square test developed by Cochran (1950), which is available in Siegel (1956). This is a generalization of the test for two related groups that has been offered by McNemar (1955, pp. 228–231). Empirical examples indicate that, when the observations take the values 1 or 0, Cochran's statistic, *Q*, gives essentially identical results with the chi square for ranked data that is related to the coefficient of concordance, provided the correction for ties is made.

One caution should be stated. As Cochran (1950) points out for his statistic, rows with identical entries (all successes or all failures) have no effect on the *Q* for columns. *Q* is equivalent to the chi square from ranks only when the computation of the latter statistic omits such rows. This omission of arrays without variance, comparable to the exclusion of ties in the sign test, may influence the experimenter's interpretation when a substantial proportion of the rows is involved.

The row means are taken as fixed in Cochran's test, and it would appear that the same holds for the chi square from ranks. For example, a different set of row means might involve a greater or smaller number of rows with identical values. In general, variances of row and column means tend to be negatively related: if the effect for objects is large, it may preclude the possibility of a significant effect for judges. Therefore, the inference from such data must be restricted to populations with distributions of row means similar to that for the data at hand.

When the experiment involves repeated judgments by each judge about each object, or when each judge judges several objects of each of several kinds, the entries can be proportions rather than 1 or 0. Under these conditions, it is possible to control column effects while testing for row effects. Mood (1950, pp. 399–402) presents two methods, an exact test for small *N*s and a test based on chi square. They require the assumption that the interactions are zero, unless either the judges or the object classes have been randomly chosen.

The Multiple Category Case

The preceding presentation has neglected the categories to which the objects belong. In most designs, there are several categories with several objects in each. The same principles of design and inference apply to this case. In a design where the same judge or judges are used for all objects, the conclusions are limited to the judge(s) as an instrument. The accuracy of judgments should be evaluated for each category separately. This can be done by obtaining a score for each case: e.g., the probability of *n* judges being correct with the chance value being determined from the relative frequency of the category in the obtained judgments. When the relative accuracy for one case can be assumed to have no effect on the accuracy for any other case, these *p* values can be pooled.

The multiple category design has one type of dependence not found in the simpler designs: the relative accuracy for one category may affect that for other categories. In the extreme case of two categories, A and B, the experimenter cannot distinguish between success in judging cases as A and success in judging cases as not-B, and conversely for B and not-A; hence no inference is possible concerning the relative success for the two categories.

Whenever the number of categories is small, some interdependence will be present and any tests of differential success for the several categories will be biased in favor of accepting the hypothesis of no difference unless this dependence is taken into account. However, an approximate test might be a one-way analysis of variance where each entry represented the difference, for a single case, between the obtained proportion and the expected proportion of correct judgments. (An index of this variety, suggested by David Wallace, was used in a complex matching study by Henry & Farley, 1959.)

Summary

This paper considers some problems of design and inference that are found in studies using the matching method with multiple judges and objects. When each object is judged by each judge, the analysis must take

into account the dependence among the several observations in each set. Failure to recognize and allow for this dependence is a common oversight in the design of matching studies. Frequently the judges or the objects must be viewed as an instrument, the conclusion being restricted to "these judges" or "these objects." Techniques are considered for testing for differences between judges, or differences between objects.

REFERENCES

Battin, I. L. On the problem of multiple matching. *Ann. math. Statist.*, 1942, **13**, 294–305.

Chapman, D. W. The statistics of the method of correct matchings. *Amer. J. Psychol.*, 1934, **46**, 287–298.

Chapman, D. W. The generalized problem of correct matchings. *Ann. math. Statist.*, 1935, **6**, 85–95.

Chapman, D. W. The significance of matchings with unequal series. *Amer. J. Psychol.*, 1936, **48**, 167–169.

Cochran, W. G. The comparison of percentages in matched samples. *Biometrika*, 1950, **37**, 256–266.

Cronbach, L. J. A validation design for qualitative studies of personality. *J. consult. Psychol.*, 1948, **12**, 365–374.

Crow, W. J. A methodological study of social perceptiveness. Unpublished doctoral dissertation, University of Colorado, 1954.

Crow, W. J. The need for representative design in studies of interpersonal perception. *J. consult. Psychol.*, 1957, **21**, 323–325.

Dudek, F. J. Determining "chance success" when a specific number of items are sorted into discrete categories. *J. consult. Psychol.*, 1952, **16**, 251–256.

Gilbert, E. J. The matching problem. *Psychometrika*, 1956, **21**, 253–266.

Hammond, K. R. Representative vs. systematic design in clinical psychology. *Psychol. Bull.*, 1954, **51**, 150–159.

Henry, W. E., & Farley, Jane. A study in validation of the Thematic Apperception Test. *J. proj. Tech.*, 1959, **23**, 273–277.

Jones, L. V., & Fiske, D. W. Models for testing the significance of combined results. *Psychol. Bull.*, 1953, **50**, 375–382.

Kelly, E. L., & Fiske, D. W. *The prediction of performance in clinical psychology.* Ann Arbor: Univer. Michigan Press, 1951.

Kendell, M. G. *Rank correlation methods.* London: Griffin, 1948.

McHugh, R. B., & Apostolakos, P. C. Methodology for the comparison of clinical with actuarial predictions. *Psychol. Bull.*, 1959, **56**, 301–308.

McNemar, Q. *Psychological statistics.* (2nd ed.) New York: Wiley, 1955.

Mood, A. M. *Introduction to the theory of statistics.* New York: McGraw-Hill, 1950.

Mosteller, F., & Bush, R. R. Selected quantitative techniques. In G. Lindzey (Ed.), *Handbook of social psychology.* Vol. I. Cambridge, Mass: Addison-Wesley, 1954. Pp. 289–334.

Roberts, A. H. Chance frequency in matching problems when success or failure is reported after each matching operation. *J. consult. Psychol.*, 1958, **22**, 233–234.

Secord, P. F. A note on the problem of homogeneity–heterogeneity in the use of the matching method in personality studies. *Psychol. Bull.*, 1952, **49**, 41–42.

Siegel, S. *Nonparametric statistics.* New York: McGraw-Hill, 1956.

Stevens, W. L. The distribution of entries in a contingency table with fixed marginal totals. *Ann. Eugen., Lond.*, 1938, **8**, 238–244.

Vernon, P. E. The evaluation of the matching method. *J. educ. Psychol.*, 1936, **27**, 1–17. (a)

Vernon, P. E. The matching method applied to investigations of personality. *Psychol. Bull.*, 1936, **33**, 149–177. (b)

Vernon, P. E. A note on the standard error in the contingency matching technique. *J. educ. Psychol.*, 1936, **27**, 704–709. (c)

Wilks, S. S. *Mathematical statistics.* Princeton Univer. Press, 1943.

(Received October 30, 1959)

Psychological Bulletin
1961, Vol. 58, No. 1, 87-88

COMMENT ON "A NOTE ON PROJECTION"

BERNARD I. MURSTEIN[1]
Interfaith Counseling Center, Portland, Oregon

In a recent issue of the *Psychological Bulletin* Chase (1960) discussed an article by Murstein and Pryer (1959). Referring to this article Chase stated that there appear to be "rather glaring faults in formulation, categorization and definition"[2] (p. 289). It would seem, however, that Chase misses the mark in two important respects:

1. There is a gross confusion in differentiating *what Pryer and I say about projection* in the critique portion of our paper from *what we report others* as saying in our review of the research.

2. There are several misinterpretations of psychological terms in Chase's criticism.

We defined "attributive" projection as "The ascribing of one's own motivations, feelings and behavior to other persons" (Murstein & Pryer, 1959, p. 354). Chase (1960), finding this definition similar to the one listed by English and English (1958), describes our definition as "clearly redundant" (p. 289). It should be noted, however, that the purpose of a review is not to invent new definitions, but to classify and order the myriad research publications bearing on the topic under consideration. The fact that our definition overlaps with one given by English and English indicates only that we have succeeded in identifying one of the common usages of the term.

We are further belabored by Chase for our putative disregard of the "unconscious" and the "self-concept" in our discussion of "attributive" projection. Again, there is a failure to understand that we did not advocate this omission but only reported a condition familiar to most persons conversant with the literature on projection—namely, that most operational usages of the term imply nothing about an unconscious or a self-concept. A typical operational definition is described by Bender and Hastorf (1953). A statement "I am wary about the trustworthiness of persons whom I do not know well" may be answered affirmatively by a subject (*S*). If he now predicts that his friend would answer the item similarly we have an example of attributive projection. The congruency of answers, however, may be based on the similarity of the personalities answering habits, experiences, or cognitive evaluations of the two *S*s without implying any notions about the unconscious or self-concept. It is exactly this lack of concern with the self-concept which led Pryer and me to criticize this approach!

Chase criticizes our use of terminology. Only the most meaningful of these comments will be considered.

1. We quoted Zilboorg's example (1935) of medieval projection which Chase regards as an example of an hallucination rather than of projection. He failed to realize that an hallucination is but one kind of pro-

[1] I should like to express my appreciation to E. Nelson Pareis and Martha Pareis for reading the manuscript and offering their valuable comments.

[2] As senior author, I assume full responsibility for all "glaring faults in formulation, categorization and definition" attributed to our paper by Chase and accordingly am replying to his comments.

jection. Thus, Murray (1933) says:

> We may speak of *perceptive projection* when sensory elements are projected, i.e., when an image takes on the vividness, substantiality, and out-thereness of a real object—as in . . . an hallucination (p. 313).

2. Chase believes that our term "rationalized projection" is simply another name for rationalization. Rationalized projection refers to an occasion in which *S* while not denying the possession of an unattractive trait or the fact that he committed some unsavory act does deny responsibility by projecting the cause of his behavior on to another. Rationalization is defined by English and English (1958) as

> the process of giving rational order or interpretation to what was previously merely a vague intuition, or was chaotic and confused (p. 438).

Since rationalization may serve to make an event comprehensible without necessarily involving recourse to self-deception, it is apparent that rationalized projection is but one of many kinds of rationalization.

3. A similar confusion of species with genus underlies Chase's attempt to equate "autistic projection" with autism. The former is a term used to describe misperceptions due to hunger, thirst, or a "set" of some kind. The latter is defined by English and English (1958) in one of their definitions as

> finding pleasure in fantasies that represent reality in wish-fulfilling terms, even when these are not believed (p. 54).

It is evident that autistic projection is one form of autistic expression.

4. In an attempt to encompass the extremely diverse and varied uses of projection, Pryer and I settled for a broad definition of it which included emotional value or need. Chase interprets this as subsuming "defecation." It is, however, unusual to regard this act as involving an emotional value for the majority of persons, Freud notwithstanding.

Finally, a new classification of types of projection is offered by Chase (1959).

> Two major categories are immediately obvious. We might term one type *defensive projection* and the other *predictive projection* (p. 289).

I should like to illustrate via a brief example why this division is not satisfactory. On the eve of election day in 1948, a noted commentator, H. V. Kaltenborn, predicted a Dewey victory. Far into the night, as the stunning reversal of expectation became manifest, Kaltenborn relied on his long experience to avoid being swayed by "early city returns" which favored Truman. Surely this is a bona fide case of predictive projection, but, does it not also smack of defensive projection?

Though I question the validity of Chase's criticisms there is no doubt that a good deal of work still remains in the area of projection. Hopefully, we will yet evolve an operational definition retaining the historical meaning, which can also be experimentally validated.

REFERENCES

BENDER, I. E., & HASTORF, A. H. On measuring generalized empathic ability (social sensitivity). *J. abnorm. soc. Psychol.*, 1953, **48**, 503–506.

CHASE, P. H. A note on projection. *Psychol. Bull.*, 1960, **57**, 289–290.

ENGLISH, H. B., & ENGLISH, AVA C. *A comprehensive dictionary of psychological and psychoanalytical terms.* New York: Longmans, Green, 1958.

MURRAY, H. A. The effect of fear upon estimates of the maliciousness of other personalities. *J. soc. Psychol.*, 1933, **4**, 310–329.

MURSTEIN, B. I., & PRYER, R. S. The concept of projection: A review. *Psychol. Bull.*, 1959, **56**, 353–374.

ZILBOORG, G. *The medical man and the witch during the Renaissance.* Baltimore: Johns Hopkins Press, 1935.

(Received August 31, 1960)

VOL. 58, NO. 2 MARCH 1961

Psychological Bulletin

VISUAL SENSITIVITY TO DIFFERENCES IN VELOCITY

ROBERT H. BROWN[1]
United States Naval Research Laboratory

In various reviews of the literature, psychologists have stressed the dependence of the perception of motion upon a multitude of factors. Kennedy (1936), for example, indicated in his review that this dependence necessitates rigorous control in the experimental method used for measuring thresholds. The need for careful analysis and experimentation, also stressed earlier by Neff (1936), has been restated more recently by Graham (1951) and Gibson (1958). Despite the caution suggested by these reviews, analysis of data available in the literature for a specific threshold proves fruitful for application to a more general form of behavior. The purpose of the present paper is to discuss this analysis.

Visual sensitivity to differences in velocity is commonly measured by presenting two objects which move at slightly different, but constant, speeds. The least detectable difference in speed is the differential threshold for the magnitude of velocity. As an initial step in the paper, consideration of angular speed indicates that it is the basic unit of measurement involved in studies of the differential threshold. Plotting differential thresholds for angular speed yields a meaningful relation to a primary variable, the speed of object motion. From these thresholds, the sensitivity is readily calculated and expressed in terms of the ratio of the threshold to the speed. As a final step in the paper, this Weber ratio for velocity is applied to tracking and other predictive behavior.

DIFFERENTIAL SPEED THRESHOLDS

Augular Speed

Graham (1951) has described the concept of visual angle and the utility of specifying stimulus extents in terms of the angle they subtend at the eye. Similarly, the visual angle per unit time or angular speed is a basic variable in experiments concerned with the visual perception of movement. Its use facilitates the comparison of data obtained under different conditions. For example, threshold measurements made in independent experiments at varying observational distances are expressed in terms of a common measure, angular speed. In addition, the use of angular speed as a stimulus specification may be necessary for good experimental design.

In Figure 1, the axis of rotation at *O* may be specified in terms of a convenient reference point such as

[1] This review of research on the visual perception of movement and its application to tracking and other predictive behavior has been improved by the suggestions of colleagues, especially Joseph Dougherty, Robert E. Gardner, Howard Gordon, Jr., and Franklin V. Taylor of the United States Naval Research Laboratory.

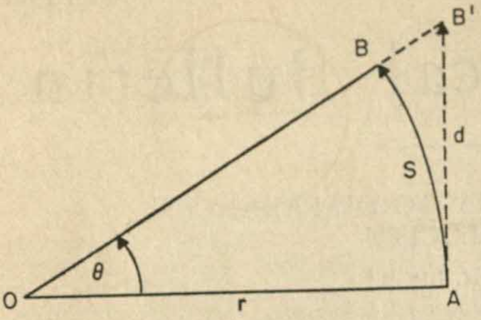


FIG. 1. Diagram representing components in the measurement of angular speed.

the front surface of the cornea. The radius of rotation (r) is given by the distance from the reference point to the appropriate moving object. When the eye looks steadily at fixation point A, the line of regard OA is stationary. Alternatively, one may assume a rotating line of regard in experiments involving fixation on a moving object. Presently available data do not indicate unequivocally that the alternative assumptions yield a measurable difference in the perception of velocity. Fleischl (1882) reported that an object seen while fixating a stationary point moves subjectively faster than when followed by the eyes. Since Aubert (1886) confirmed the phenomenon, it has been called the Aubert-Fleischl paradox. However, the need for reexamination of the paradox is indicated by the recent work of Gibson, Smith, Steinschneider, and Johnson (1957). When they measured the accuracy of visual perception of motion, they found no difference for the two modes of observation.

As a stimulus rotates about the reference point in Figure 1, its instantaneous angular speed (ω) is given by:

$$\omega=\frac{d\theta}{dt} \qquad [1]$$

where θ is the angle swept by the radius vector r in time t. The value of θ is given by:

$$\theta=\frac{s}{r} \quad \text{(in radians)} \qquad [2]$$

or

$$\theta=\frac{57.3s}{r} \quad \text{(in degrees)} \qquad [3]$$

The measure *angular speed* may be used advantageously not only for rotational motion but also for tangential motion. In Figure 1, the rectilinear distance d is a close approximation to the arc s for angular displacements of the magnitude usually used. For example, d exceeds s by only 1% for a θ of 10°. Conversely, angular displacements less than 10° may be calculated with less than 1% error by substituting d for s in Equation 3. For greater displacements, θ is calculated from:

$$\theta=\arctan\frac{d}{r} \qquad [4]$$

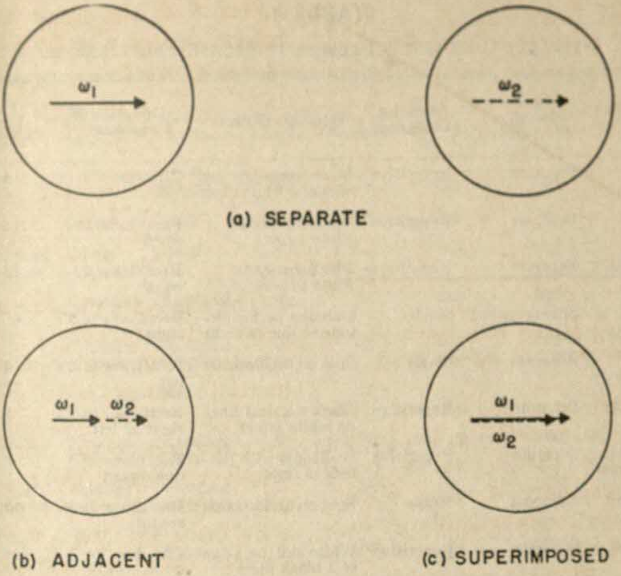
For uniform angular motion when ω is constant:

$$\omega=\frac{\theta}{t} \qquad [5]$$

Although this equation is a special case of the earlier definition of instantaneous angular speed in derivative form, it applies with very few exceptions to experiments which have been conducted on the perception of movement. By substitution for θ from Equation 2, uniform angular speed may be described by:

$$\omega=\frac{s}{rt} \quad \text{(in radians per unit time)} \qquad [6]$$

where the arc s and the radius r are expressed in the same units. As an approximation for small angular dis-



(a) SEPARATE

(b) ADJACENT

(c) SUPERIMPOSED

FIG. 2. Procedures used in presentations of stimulus motion.

placements, we may substitute *d* for *s* to obtain:

$$\omega=\frac{d}{rt}=\frac{v}{r}$$

(in radians per unit time) [7]

or

$$\omega=\frac{57.3v}{r}$$

(in degrees per unit time) [8]

where the uniform linear speed *v* and the observational distance *r* are expressed in consistent units.

The Differential Speed Threshold and Its Measurement

The differential threshold for angular speed, $\Delta\omega$, may be defined in terms of:

$$\Delta\omega=\omega_2-\omega_1 \qquad [9]$$

where ω_2 is a uniform angular speed an observer discriminates according to a specified criterion from the constant rate of motion ω_1. In measurements of $\Delta\omega$, the spatial relationship of ω_1 and ω_2 is critical. Three procedures used to date involve stimuli which are separate, adjacent, and superimposed. In Figure 2, a circle represents schematically an outline of a display, such as moving belt, rotating disc, or cathode-ray tube, used in presenting ω_1 and ω_2. The speeds are represented by the vectors in each display. In Procedure *a*, the stimuli for the two speeds are spatially apart and are viewed by looking from one display to the other. In Procedure *b*, the stimuli are in immediate proximity. In Procedure *c*, they are superimposed on each other. Table 1 summarizes the most significant stimulus conditions present in measurements of $\Delta\omega$.

At least six experiments have been reported for measurements involving separate stimuli. Bourdon (1902) utilized two rotating white discs with a black rectangle on the edge of each. The subject adjusted the speed of one in increments until it was noticeably faster than the other. Similar meas-

TABLE 1

STIMULUS CONDITIONS PRESENT IN MEASUREMENTS OF $\Delta\omega$

Experimenters	Spatial Relation of ω_1 and ω_2	Stimulus Frequency	Stimulus Objects	Direction of Motion	Field Extent (degrees)	Observational Distance (cm.)
Bourdon (1902)	Separate	Repetitive	Black rectangle on edge of 2 white discs	Circular	6.4	200
Brown (1931)	Separate	Repetitive	Black square on white paper	Rectilinear upward	2.15–4.30	200
Brown & Mize (1932)	Separate	Repetitive	Black square on white paper	Rectilinear upward	2.15–4.30	200
Zegers (1948)	Superimposed	Single	2 needles perpendicular to line of sight	Rectilinear to *S*'s right	3.6 –15.0	15.9
Hick (1950)	Adjacent	Single	Spot on oscilloscope	Rectilinear to *S*'s left	4.8	53.3
Ekman & Dahlbäck (1956)	Separate	Repetitive	Black vertical lines on white paper	Rectilinear to *S*'s right or left	5.72	50
Gibson, Smith, Steinschneider, & Johnson (1957)	Separate	Repetitive	Wallpaper with pattern of dots	Rectilinear downward	8.4	122
Notterman & Page (1957)	Adjacent	Single	Spot on oscilloscope	Rectilinear horizontal	10.0	25.4
Brandalise & Gottsdanker (1959)	Separate	Repetitive	White dot on edge of 2 black discs	Circular	5.2	200

urements were made by Brown (1931) and by Brown and Mize (1932) for a black square moving upward on white paper which the observer saw in either of two windows. Ekman and Dahlbäck (1956) and Gibson et al. (1957) have made measurements involving the adjustment of ω_2 for apparent equality with ω_1. The former utilized two apertures in each of which alternately the observer saw the horizontal motion of black vertical lines on white paper. The latter presented behind two windows a downward moving wallpaper with a pattern of dots. Most recently, Brandalise and Gottsdanker (1959) have had subjects adjust the speed of rotation of a black disc with a white dot on its edge to apparent equality with that of another. In these six experiments, the measurements of $\Delta\omega$ were based on comparisons of the two speeds which were viewed separately in different places. Since the equipment involved rotating drums or discs, stimulation was repetitive.

Use of a moving spot on an oscilloscope has facilitated presentation of adjacent stimuli. During rectilinear motion of a pip at constant speed, an incremental change in speed is introduced. Hick (1950) and Notterman and Page (1957) measured the differential threshold in speed for a pip as it was horizontally deflected across the face of a cathode-ray tube. Temporal features of this procedure differ from the first. The stimuli are presented only once and then in immediate succession.

The procedure of superimposed stimuli may be illustrated by monocular movement parallax. When two objects move at the same linear speed perpendicular to the subject's line of sight, the difference in their angular speeds provides an indication of their distances from the subject. As the objects are brought closer together, the difference in angular speeds decreases to a threshold value. Zegers (1948) has measured the differential threshold speed for two needles by

TABLE 2

METHODOLOGY USED IN THE MEASUREMENTS OF $\Delta\omega$

Experimenters	Psychophysical Method	Measure of $\Delta\omega$	No. of Speeds	No. of Subjects	No. of Measurements per Speed per Subject	Total No. of Measurements	Speed (degrees per sec.)	
							Minimum	Maximum
Bourdon (1902)	Limits	Mean	3	1	20	60	0.77	5.04
Brown (1931)	Limits	Mean	2	2	10	40	1.79	3.58
Brown & Mize (1932)	Limits	Mean	6	2–3	3–6	117	1.72	4.58
Zegers (1948) 3.6° field	Constant stimuli	Standard deviation	4	2	100	800	2.67	20.1
15.0° field	Constant stimuli	Standard deviation	6	2	100	1200	2.67	36.1
Hick (1950)	Constant stimuli	Mean	7	18	—	—	0.15	10.2
Ekman & Dahlbäck (1956)	Average error	Standard deviation	5	10	4	200	2.07	4.81
Gibson, Smith, Steinschneider, & Johnson (1957)	Average error	Standard deviation	1	24	10	240	4.80	4.80
Notterman & Page (1957)	Constant stimuli	Mean	7	10	30	2100	0.34	22.7
Brandalise & Gottsdanker (1959)	Average error	Standard deviation	5	10	50	2500	2.7	24.3

this procedure, which temporally involves the single presentation simultaneously of ω_1 and ω_2.

The psychophysical method used has been less critical for measurements of $\Delta\omega$ than the spatial relationship of the stimuli. Table 2 lists significant methodological characteristics for the nine experiments. Items specially worth noting are the limited range of speeds in most experiments and the small number of measurements in some studies.

The Differential Threshold as a Function of Speed

The marked effect of spatial order may be observed by inspection of Figure 3, in which $\Delta\omega$ is plotted against ω. The curves and their points represent the use of adjacent, separate, and superimposed stimuli. The measure of $\Delta\omega$ is that indicated in Table 2. Since Brown (1931) and Brown and Mize (1932) made only a small number of exploratory measurements, the points plotted for their experiments are the geometric means of values they reported for speeds 1–2, 2–3, 3–4, and 4–5° per second. The data plotted for superimposed stimuli represent the monocular movement parallax thresholds obtained by Zegers (1948) with the widest and narrowest visual fields of the four for which he made measurements. Otherwise, the points represent all values reported in the literature for $\Delta\omega$ as listed in Table 2.

The solid lines have been drawn with unit slope and represent a constant Weber fraction ($\Delta\omega/\omega$). In the case of adjacent stimuli, solution for the intercept constant by the method of least squares yields the plotted equation:

$$\log \Delta\omega = -0.859 + \log \omega \quad [10]$$

It may be observed as a rough approximation that the differential threshold increases in direct proportion to the angular speed of a stimu-

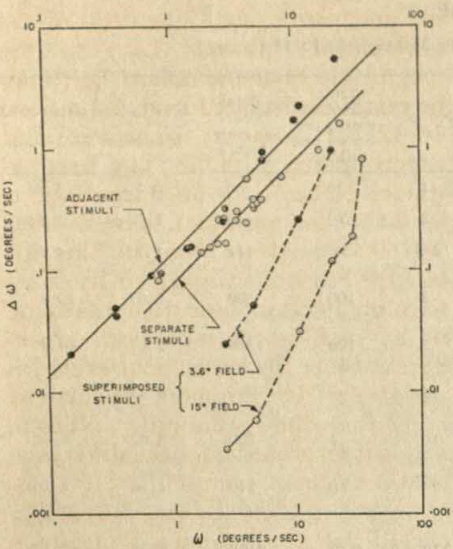


Fig. 3. The differential angular speed threshold ($\Delta\omega$) as a function of the angular speed (ω) of stimuli which are presented spatially adjacent, separate, and superimposed.

lus. Discrepancies in this approximation occur in a middle range of speeds (1–5° per second) where the measured $\Delta\omega$ falls below the empirical straight line. At faster speeds, $\Delta\omega$ increases at an increasing rate with ω.

For separate stimuli, the least squares equation is as follows:

$$\log \Delta\omega = 1.114 + \log \omega \qquad [11]$$

Under these conditions, the differential threshold increases in direct proportion to speed from approximately 1 to 10° per second. The differential threshold is greater at slow speeds, and less at fast speeds, than the best fitting straight line of unit slope.

Data obtained for superimposed stimuli can be described by a constant Weber fraction only under quite restricted conditions. Thus, for the widest field (15°) a solid line is drawn between the points for the two slowest speeds. Its equation is:

$$\log \Delta\omega = -2.893 + \log \omega \qquad [12]$$

The rapid increase in the differential threshold with speed for superimposed stimuli may be interpreted in terms of instability of the retinal image and intensity effects in individual cones.

As Zegers (1948) indicates in discussion of his results, high speeds interfere with good "pickup" of the stimuli as they appear in the visual field and, also, with adequate following movements of the eyes. The influence of extent of visual field, so marked in Figure 3, was markedly decreased, if not eliminated, by providing appropriate aids during control experiments to fixation and stimulus "following." Careful measurement of the vertical distance between the curves for the 15 and 3.6° fields indicates that they could very nearly be superimposed by a shift of 0.905 log unit, the mean of separations of 0.853, 0.947, 0.909, and 0.909 log unit. We may infer that the vertical position of the curves depends primarily on stability of the retinal image. When stimulus conditions for good fixation of the stimulus are absent, the differential threshold function of Figure 3 is shifted uniformly upward with decrease in extent of the visual field.

The shape of the curves for superimposed stimuli appears to be dependent upon the intensity effects occurring in individual cones. Evidence for this inference is less direct than Zegers' control experiments involving improved conditions for fixation and pursuit of the stimulus. However, it should be pointed out that Graham, Baker, Hecht, and Lloyd (1948) measured the differential threshold as a function of the luminance of the stimulus field. Neutral tint filters were placed behind the metal tube through which the observer saw two needles, one above the other, moving at constant and equal speeds back and forth across an illuminated field. Measurements of the precision of distance settings with one needle yielded differential angular speeds for different luminances of the visual field. The decrease in $\Delta\omega$ as a function of the increase in luminance is described by Hecht's intensity discrimination equation upon the assumption that $\Delta\omega$ is a measure of differences in diffraction luminances and provides a ΔI seen against the general illumination, I.

In addition, measurements of the threshold luminance for a moving spot of light indicate that the intensity effect of speed is similar to the parallax effect of Figure 3. At moderate speeds, the threshold luminance for discrimination of motion increases in direct proportion to the speed (Brown, 1958). At faster speeds (greater than 10° per second), the luminance threshold increases at a disproportionate rate until it approximates an asymptote at a limiting speed (30 to 40° per second). This relationship, like that found by Zegers, may be interpreted in terms of intensity effects occurring in individual cones. As angular speed increases, the duration of passage of the image across a given cone is shortened. Since the intensity effect in each receptor unit is lessened, the luminance for the moving spot or the differential angular speed of the needles must be increased.

The Weber Ratio

The Weber ratio provides a convenient measure by means of which velocity discriminations may be compared with other sensory discriminations and with performance in tracking and predicting. The ratio of the differential threshold ($\Delta\omega$) to the magnitude of the standard (ω_1) may be readily calculated from Equations 10–12 for adjacent, separate, and superimposed stimuli. The best estimate of $\Delta\omega/\omega$ for an unspecified ω is 0.138 for adjacent stimuli and 0.0769 for separate stimuli. This difference has been confirmed by Notterman (1959) in measurements made by an oscilloscope with both prececdures. Since his experiment excludes variations in stimulus conditions other than the spatial order, Notterman's interpretation of the difference is of particular interest:

> Subjects in the adjacent presentation case can base their discrimination on a comparison of the amount of time taken to traverse the initial and final 1½ inches on the scope face, or—and this is important—they can disregard time and look for the jerk which occurs when the moving spot instantaneously increases its velocity. The subjects employing the separate presentation procedure do not have this option: since the standard and comparison stimuli are separated in time, there is no jerk. In short, the subjects of the (adjacent) procedure may have changed the problem from one requiring a comparison of two velocities, to one requiring a judgment of the presence or absence of jerk (p. 3).

The marked superiority of superimposed stimuli in yielding a low Weber fraction is illustrated by the value of 0.00128 for two needles traversing an extent of 15° at angular speeds less than 5° per second. This superiority is readily understandable. Superimposition of one needle in front of the other provides an angular offset which Zegers has found to be a basic determiner of the differential angular speed threshold. The angular offset is absent when stimuli are presented adjacently in immediate succession or separately in space and time.

Variation of the Weber fraction over the whole speed range is plotted

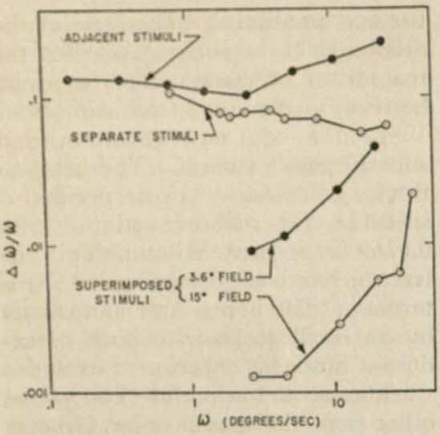


FIG. 4. The Weber ratio ($\Delta\omega/\omega$) as a function of the angular speed (ω) for discriminations utilizing adjacent, separate, and superimposed stimuli.

in Figure 4. The points represent geometric means of values determined by different investigators at approximately the same angular speeds. Thus, the top curve for adjacent stimuli is the average of values obtained by Hick (1950) and Notterman and Page (1957). Except for the point at the slowest speed (Hick) and at the fastest speed (Notterman and Page), each point is the geometric mean of the Weber fraction in both studies. A similar procedure has been followed in averaging measurements made with separate stimuli. For superimposed stimuli, the Weber fraction has been calculated directly from Zegers' data. In this case, the ratio is directly proportional to the angular offset existing between the reference and comparison stimuli. As Zegers has indicated, the value of the angular offset (and the Weber ratio) increases with speed.

Examination of the curves of Figure 4 suggests a useful empirical generalization. The Weber fraction for nonsuperimposed stimuli is approximately constant in the middle range of angular speeds. Thus, in the range of 0.1 to 20° per second, $\Delta\omega/\omega$ shows no greater change than a doubling. For adjacent stimuli, the maximal ratio is only 2.2 times greater than the minimal ratio. For separate stimuli, there is a change by a factor of 1.9. Although the Weber fraction may be fairly constant in the middle range of stimulus values, the rapid rise of the curve for superimposed stimuli suggests that the ratio may increase markedly at extremes.

The constancy of the Weber ratio for differential speed thresholds may be interpreted at a descriptive level for comparison with other sensory discriminations. Woodworth and Schlosberg (1954) have indicated that for many sensory discriminations the differential threshold is a measure of the variability of the effects of stimulation, i.e., $\Delta\omega = K\sigma_{\omega}$. For discriminations of motion according to Brown (1960), the variability in turn is proportional to the speed, i.e., $\sigma_{\omega} = C\omega$. It is therefore not surprising that $\Delta\omega/\omega$ is constant, at least within limits which are not too well defined in Figure 4.

It is of interest to compare the magnitude of the ratio with that for other discriminations. Under optimal conditions, the minimal Weber fraction with superimposed stimuli is comparable to that measured for pitch discrimination with a standard tone and a comparison tone differing slightly in frequency. Measurements of pitch discrimination indicate that Weber's fraction is constant at about 0.002 beyond 250 cycles per second, rising somewhat at the lower frequencies. The differential speed threshold ratio, as measured with separate stimuli, is comparable to the Weber fraction for lifted weights. When measured by lifting weights successively with one hand, the Weber ratio is approximately 0.075 for weights greater than 200 grams. We may conclude that the Weber ratio for differential speed thresholds not only is constant in a medium range of stimulus values, but also is of the same order of magnitude as that found for other discriminations.

Tracking Behavior

Studies of tracking behavior illustrate an application of the Weber ratio for differential speed thresholds. This application is of particular interest since earlier reviews have emphasized the significant motor characteristics of tracking behavior (Birmingham & Taylor, 1954; Fitts, 1951). Perceptual characteristics have been implied by occasional observations that an operator tracks a target quickly and efficiently under optimal conditions because he estimates its present speed and acceleration and thereby anticipates its future motion. During World War II, for example, the systematic investigation of the manual controls for antiaircraft fire control systems indicated the anticipatory nature of tracking, as discussed by Helson (1949).

Foxboro studies. In the Foxboro studies directed by Helson, error was recorded for compensatory tracking in which the tracker tries to keep a moving pointer aligned as much as possible with a stationary reference pointer. Compensatory tracking may be contrasted with pursuit tracking in which both pointers move and the tracker aligns the following cursor under his control with the moving target pointer. In the Foxboro studies the tracker compensated for the displacement of a moving pointer, representing the aiming point, from the actual position indicated by a stationary pointer. Tracking error was measured by the time required for the target to move from its actual position to the aiming point.

Speed of the handwheel rotation was a major variable controlling tracking accuracy. For a constant-speed unidirectional course, with an increase in rate of cranking, the tracking error decreased from 55 to 6 milliseconds when a light handwheel of 2.25-inch radius was used (Foxboro Company, 1943a). Since the tracking error was consistently of the order of milliseconds and could be as small as one hundredth of the fastest reaction time, it is evident that the tracker anticipated the future motion of the target and thereby avoided the series of oscillations his long reaction time would otherwise produce.

For simple sinusoidal courses, the tracker not only anticipated the motion of the target but also used an averaging motion of the handwheel when the course was of too high a frequency to follow exactly. As course frequency increased, the tracker eliminated terminal portions of swings. Inertia in the form of a heavy handwheel or a flywheel effect smoothed the direct tracking of courses not requiring high accelerations and rapid reversals in direction (Foxboro Company, 1943b). In addition, the averaging type of behavior was dependent upon practice and familiarity with the course being tracked.

Contemporary models for tracking behavior. Since World War II, the concept of feedback mechanisms has been generalized to the entire field of control and communication theory in machines and animals (Wiener, 1948). As applied to antiaircraft fire control behavior, the concept states that the tracker uses the difference between the stimulus of a target's motion and his response as a new input to make his motion correspond more closely to that of the target. Engineers analyzed human tracking performance in terms of simple servo systems with feedback (James, Nichols, & Phillips, 1947; Raggazini, 1948; Tustin, 1947). Stimulated by the mathematical systems equations which emerged from this analysis, psychologists have developed their own models to describe the behavior involved in minimizing the difference between two positions with control of one (Birmingham & Taylor, 1954; Fitts, 1951; Noble, Fitts, & Warren, 1955). These models make two basic assumptions: intermittency of response, and predictiveness of response.

Intermittency of tracking responses. Despite the smooth and apparently continuous appearance of efficient tracking, experimental evidence from several sources indicates that the tracker responds intermittently. First, a time record of tracking performance shows a typical periodicity with a predominant frequency of two responses per second (Craik, 1947; Ellson, Hill, & Gray, 1947). Second, analysis of the response patterns to a step input displacement of position shows that quick corrective movements occur without visual or kinesthetic guidance and that the typical time for completing a corrective movement, including reaction time, is approximately 0.5 second (Chernikoff & Taylor, 1952; Searle & Taylor, 1948; Taylor & Birmingham, 1948). Third, the assumption that the tracker responds intermittently at 0.5-second intervals during continuous tracking agrees with the optimal time constant obtained for conventional aided tracking (Birmingham & Taylor, 1954; Mechler, Russell, & Preston, 1949). Fourth, with the assumption of 0.5-second intermittency of corrections, one may predict the optimal time constants for more complex aided-tracking control systems involving an acceleration component as well as the conventional position and rate controls (Searle, 1951).

Predictiveness of tracking responses. The assumption of predictiveness in tracking responses is supported by the following findings. First, the Foxboro studies showed that the time error for manual handwheel tracking is much less than the reaction time, as discussed above. Second, pursuit tracking usually yields lower error scores than compensatory tracking (Chernikoff, Birmingham, & Taylor, 1955; Poulton, 1952; Senders & Cruzen, 1952). In the pursuit mode of tracking, responses may be made on the basis of a predictable course of the target since its marker moves independently of the marker with which the tracker follows. In the compensatory mode, prediction must be limited to the tracking error since the tracker attempts to stabilize a moving marker representing the difference between target motion and his own control motion. Third, Chernikoff et al. (1955) found that an aided-tracking control impairs performance for the pursuit mode but materially improves it for the compensatory situation. They resolved this apparently paradoxical finding by considering the nature of aided-tracking controls in terms of the predictiveness of tracking responses. With a position control, the position of the moving marker controlled by the tracker is directly proportional to the position of his control. With aided tracking, a movement of the control not only causes a proportional change in the position of the marker, but also introduces a change in its rate of motion. The aided-tracking time constant is yielded by the ratio of the control sensitivities. With the proper time constant in compensatory tracking, the operator can correct an error with a control motion proportional to the position component of the error. He thereby sets in changes in rate of motion in amounts that are correct on the average to match the target motion. Use of the aided control in pursuit tracking requires that the tracker ignore target velocity and not attempt to predict future position. Later experiments by Chernikoff and Taylor (1957) have indicated an effect of target speed on the optimal time constant for both pursuit and compensatory tracking.

Tracking error. With verification of the assumptions of intermittency and predictiveness for tracking performance, it is evident how differential speed thresholds limit the tracker's responses with a position control. It may be assumed that at a given instant in time the tracker is exactly on target but that his cursor and the target are moving at different speeds. During a short period of time, the position error generated is approximately the product of this speed difference and the temporal interval. Since response intermittency holds the temporal interval constant, the tracking error is directly proportional to the speed difference which the tracker can discriminate.

Speed of target motion seems to have the same effect on tracking error as it has on the differential speed threshold as measured with nonsuperimposed stimuli, i.e., tracking error increases as a linear function of speed. Bowen and Chernikoff (1958) have investigated the relationship between magnification, speed of target motion, and tracking error with a compensatory position-control system. Both with and without magnification, measures of tracking performance did not vary for a constant target speed when the frequency and amplitude of motion were varied over a range useful in tracking research. Tracking error increased with an increase in average speed of target motion. Departures from a linear relationship were not large.

Predictive Behavior

Prediction Motion

Data from Gottsdanker's series of studies of prediction motion demonstrate a marked similarity of pursuit tracking error to the differential speed threshold for adjacent stimuli. Similar to the differential threshold (approximately 14% of the speed) is the average error a tracker makes in following a target which moves at a constant speed but suddenly disappears. During the second following the disappearance, the tracker maintains the speed with an average error of 13, 14, and 16%, as measured in three separate studies by Gottsdanker (1952a, 1952b, 1955).

On some trials when the target was accelerating or decelerating at the moment of disappearance, the tracker did not continue the uniform change in speed. It should be noted, however, that at the moment of disappearance the change in speed for a 0.5-second interval was only 5 to 7% of the speed and presumably was below the tracker's threshold. Gottsdanker (1956) has reviewed the experimental literature on responses to acceleration of target motion. He concluded that smoothly accelerated motion is generally responded to as if the speed were constant, i.e., the change in speed did not exceed the differential speed threshold in the studies cited.

Gottsdanker (1952a) has measured the tracking error not only for disappearing targets, but also for completed courses. The measured error is consistent with one calculated upon the basis of the assumptions of a 0.5-second intermittency in response and a 14% speed threshold. The average error in tracking a target moving at a constant speed of 8 millimeters per second was 0.50 millimeters. If the tracker were exactly on target at a given instant, his error a half-second later would be calculated from the assumptions as the product of 0.14×8×0.50 or 0.56 millimeters, and the average error during the interval should be 0.28 millimeter. It may be assumed more realistically that the tracker was not exactly on target at the beginning of the interval. The average error should be calculated as correspondingly greater than the minimal value of 0.28 millimeter.

The prediction of tracking error from the Weber ratio for speed discriminations is not limited to visually presented stimuli, but may be extended to other stimuli. Gottsdanker (1954) has measured the precision of tapping at a constant rate of two per second. He found that subjects could maintain this rate to an accuracy of 2.4% when the stimulus of pops from a magnetic tape playback was removed. In the Foxboro studies it was found that the tracker could utilize the increased precision of rapid repetitive movements in fast handwheel cranking over the intermittent corrective responses of slower handwheel turning. As an approximation, the tracking error should be limited by the product of the repetition rate threshold and the time for each repetitive movement at the faster speeds. For example, the time error should be the product of 0.024 and 0.25 second per repetitive movement for cranking at 120 rpm. This value coincides exactly with the measured time error of 6 milliseconds for the light handwheel with short radius.

Prediction of Future Positions of a Moving Target

Although the differential speed threshold would seem to be clearly related to predictions of future position of a moving object, data on the nature of the relationship are limited. Slater-Hammel (1955) has had subjects observe a marker moving at a uniform speed over different display distances and then had them estimate when the marker would complete traversing different target distances. The display distance did not affect the error in time of estimating the arrival of the uniformly moving marker at a specified point in space. However, the error increased systematically with an increase in the target distance which the marker traversed after disappearing. In terms of percentage of the required time, the error varied between 8.9% and 21.6%. These values agree with expectations based on the Weber ratio for speed discriminations with nonsuperimposed stimuli (cf. Figure 4).

Morin, Grant, and Nystrom (1956) have reported similar results despite two important differences in their experimental procedure. First, instead of Slater-Hammel's stimulus which moved continuously at a constant speed, Morin et al. used the successive illumination of cue lights which were placed at even intervals in a horizontal row. After illumination of the last cue light, the subject estimated the time it would take the imaginary moving object to reach a target light. Second, the object traveled at a rather slow computed speed of either 0.179 or 0.358° per second rather than the speed of approximately 5° per second used by Slater-Hammel. Results obtained by Morin et al. confirmed the fact that the error of estimating arrival increases with target distance. Significantly, they also found for their faster speed that the mean errors of estimation were generally less than 10% of the computed time. When the speed was 0.179° per second, the mean errors of estimation ranged from 25 to 53%. These values suggest an apparent extrapolation to slow speeds of the data presented in Figure 4.

Garvey, Knowles, and Newlin (1956) have measured the accuracy of prediction in terms of deviations in range and bearing between estimated and actual position plots on four different radar displays. They found that accuracy of estimated position was a function of target speed, i.e., the faster the motion of the target the less accurate the estimate. This relationship resembles that between $\Delta\omega$ and ω of Figure 3.

Gottsdanker and Edwards (1957) have studied a more complex type of prediction situation. Two targets moved down perpendicular paths towards an intersection but disappeared before reaching it. The subject estimated where one target would be when the other crossed the intersection. Gottsdanker concluded that for both accelerated and constant-speed targets the prediction was based on relative positions at time of the target's disappearance rather than on relative speeds or accelerations.

Summary

Measurements of the differential speed threshold ($\Delta\omega$) have been plotted against speed (ω) for comparison stimuli which were presented adjacent, separate, or superimposed. As a rough approximation, the threshold increases in direct proportion to speed for nonsuperimposed stimuli over a range from 0.1 to 20° per second ($\Delta\omega = K\omega$). Although the relationship for superimposed stimuli (monocular parallax) is similar, inadequate ocular following movements and receptor intensity effects modify the relationship at fast speeds (greater than 5° per second). Estimates of the Weber ratio ($\Delta\omega/\omega$) of 0.138 for adjacent stimuli and of 0.0768 for separate stimuli provide a basis for interpretation of tracking and other predictive behavior. Experiments support the assumptions of intermittency and predictiveness of responses in tracking. With these assumptions, error in performance may be calculated for relatively simple tasks from the Weber ratio. For more complex tasks, constancy of the Weber ratio agrees with the linear relationship found between tracking error and speed of target motion.

References

Aubert, H., Die Bewegungsempfindung. *Arch. ges. Physiol.*, 1886, **39**, 347–370.

Birmingham, H. P., & Taylor, F. V. A human engineering approach to the design of man-operated continuous control systems. *USN Res. Lab. Rep.*, 1954, No. 4333.

Bourdon, B. *La perception visuelle de l'espace.* Paris: Reinwald, 1902.

Bowen, J. H., & Chernikoff, R. The effects of magnification and average course velocity on compensatory tracking. *USN Res. Lab. Rep.*, 1958, No. 5186.

Brandalise, B. B., & Gottsdanker, R. M. The difference threshold of the magnitude of visual velocity. *J. exp. Psychol.*, 1959, **57**, 83–88.

BROWN, J. F. The thresholds for visual movement. *Psychol. Forsch.*, 1931, **14**, 249–268.

BROWN, J. F., & MIZE, R. H. On the effect of field structure on differential sensitivity. *Psychol. Forsch.*, 1932, **16**, 355–372.

BROWN, R. H. Influence of stimulus luminance upon the upper speed threshold for the visual discrimination of movement. *J. Opt. Soc. Amer.*, 1958, **48**, 125–128.

BROWN, R. H. Some methodological considerations in measuring visual thresholds to velocity. *Percept. mot. Skills*, 1960, **11**, 111–122.

CHERNIKOFF, R., BIRMINGHAM, H. P., & TAYLOR, F. V. A comparison of pursuit and compensatory tracking under conditions of aiding and no aiding. *J. exp. Psychol.*, 1955, **49**, 55–59.

CHERNIKOFF, R., & TAYLOR, F. V. Reaction time to kinesthetic stimulation resulting from sudden arm displacement. *J. exp. Psychol.*, 1952, **43**, 1–8.

CHERNIKOFF, R., & TAYLOR, F. V. Effects of course frequency and aided time constant on pursuit and compensatory tracking. *J. exp. Psychol.*, 1957, **53**, 285–292.

CRAIK, K. J. W. Theory of the human operator in control systems: I. The operator as an engineering system. *Brit. J. Psychol.*, 1947, **38**, 56–61.

EKMAN, F., & DAHLBÄCK, B. A subjective scale of velocity. Report No. 31, 1956, Psychological Laboratory, University of Stockholm.

ELLSON, D. G., HILL, H., & GRAY, F. Wave length and amplitude characteristics of tracking error curves. *USAF Air Material Command memo. Rep.*, 1947, No. TSEAA-694–2D.

FITTS, P. M. Engineering psychology and equipment design. In S. S. Stevens (Ed.), *Handbook of experimental psychology*. New York: Wiley, 1951. Pp. 1287–1340.

FLEISCHL, E. Physiologische-optische notizen. *SB Akad. Wiss. Wien*, 1882, **86**, 17–25.

FOXBORO, COMPANY. Handwheel speed and accuracy of tracking. *OSRD Rep.*, 1943, No. 3453. (Foxboro Co., PB 40615) (a)

FOXBORO COMPANY. Inertia, friction, and diameter in handwheel tracking. *OSRD Rep.*, 1943, No. 3454. (Foxboro Co., PB 40614) (b)

GARVEY, W. D., KNOWLES, W. B., & NEWLIN, E. P. Prediction of future position of a target track on four types of displays. *USN Res. Lab. Rep.* 1956, No. 4721.

GIBSON, J. J. Research on the visual perception of motion and change. In *Second symposium on physiological psychology*. Washington, D. C. ONR, 1958. Pp. 165–176.

GIBSON, J. J., SMITH, O. W., STEINSCHNEIDER, A., & JOHNSON, C. W. The relative accuracy of visual perception of motion during fixation and pursuit. *Amer. J. Psychol.*, 1957, **70**, 64–68.

GOTTSDANKER, R. M. The accuracy of prediction motion. *J. exp. Psychol.*, 1952, **43**, 26–36. (a)

GOTTSDANKER, R. M. Prediction-motion with and without vision. *Amer. J. Psychol.*, 1952, **65**, 533–543. (b)

GOTTSDANKER, R. M. The continuation of tapping sequences. *J. Psychol.*, 1954, **37**, 123–132.

GOTTSDANKER, R. M. A further study of prediction-motion. *Amer. J. Psychol.*, 1955, **68**, 432–437.

GOTTSDANKER, R. M. The ability of human operators to detect acceleration of target motion. *Psychol. Bull.*, 1956, **53**, 477–487.

GOTTSDANKER, R. M., & EDWARDS, R. V. The prediction of collision. *Amer. J. Psychol.*, 1957, **70**, 110–113.

GRAHAM, C. H. Visual perception. In S. S. Stevens (Ed.), *Handbook of experimental psychology*. New York: Wiley, 1951. Pp. 868–920.

GRAHAM, C. H., BAKER, K. E., HECHT, M., & LLOYD, V. V. Factors influencing thresholds for monocular movement parallax. *J. exp. Psychol.*, 1948, **38**, 205–223.

HELSON, H. Design of equipment and optimal human operation. *Amer. J. Psychol.*, 1949, **62**, 473–497.

HICK, W. E. The threshold for sudden changes in the velocity of a seen object. *Quart. J. exp. Psychol.*, 1950, **2**, 33–41.

JAMES, H. M., NICHOLS, N. B., & PHILLIPS, R. S. *Theory of servomechanisms*. New York: McGraw-Hill, 1947.

KENNEDY, J. L. The nature and physiological basis of visual movement discrimination in animals. *Psychol. Rev.*, 1936, **43**, 494–521.

MECHLER, E. A., RUSSELL, J. B., & PRESTON, M. G. The basis for the optimum aided-tracking time constant. *J. Franklin Inst.*, 1949, **248**, 327–334.

MORIN, R. E., GRANT, D. A., & NYSTROM, C. O. Temporal predictions of motion inferred from intermittently viewed light stimuli. *J. gen. Psychol.*, 1956, **55**, 59–71.

NEFF, W. S. A critical investigation of the visual apprehension of movement. *Amer. J. Psychol.*, 1936, **48**, 1–42.

NOBLE, M., FITTS, P. M., & WARREN, C. E. The frequency response of skilled subjects in a pursuit tracking task. *J. exp. Psychol.*, 1955, **49**, 249–256.

NOTTERMAN, J. M. Research on visual discrimination of velocity. Status Report

No. 3, 1959, Princeton University, USAF Office of Scientific Research, Contract AF 49 (638)-381.

NOTTERMAN, J. M., & PAGE, D. E. Weber's law and the difference threshold for the velocity of a seen object. *Science*, 1957, **126**, 652.

POULTON, E. C. Perceptual anticipation in tracking with two-pointer and one-pointer displays. *Brit. J. Psychol.*, 1952, **43**, 222–229.

RAGGAZINI, J. R. Engineering aspects of the human being as a servomechanism. Paper read at American Psychological Association, Boston, September 1948.

SEARLE, L. V. Psychological studies of tracking behavior: IV. The intermittency hypothesis as a basis for predicting optimum aided-tracking time constants. *USN Res. Lab. Rep.*, 1951, No. 3872.

SEARLE, L. V., & TAYLOR, F. V. Studies of tracking behavior: I. Rate and time characteristics of simple corrective movements. *J. exp. Psychol.*, 1948, **38**, 615–631.

SENDERS, J. W., & CRUZEN, M. Tracking performance on combined compensatory and pursuit tasks. *USAF WADC tech. Rep.*, 1952, No. 52–39.

SLATER-HAMMEL, A. T. Estimation of movement as a function of the distance of movement perception and target distance. *Percept. mot. Skills*, 1955, **5**, 201–204.

TAYLOR, F. V., & BIRMINGHAM, H. P. Studies of tracking behavior: II. The acceleration pattern of quick manual corrective responses, *J. exp. Psychol.*, 1948, **38**, 783–795.

TUSTIN, A. The nature of the operator's response in manual control and its implications for controller design. *J. Inst. Elec. Engr.*, 1947, **94**, 190–202.

WIENER, N. *Cybernetics or control and communication in the animal and the machine.* New York: Wiley, 1948.

WOODWORTH, R. S., & SCHLOSBERG, H. *Experimental psychology.* New York: Holt, 1954.

ZEGERS, R. T. Monocular movement parallax thresholds as functions of field size, field position, and speed of stimulus movement. *J. Psychol.*, 1948, **26**, 477–498.

(Received January 21, 1960)

Psychological Bulletin
1961, Vol. 58, No. 2, 104–121

SELF-ACCEPTANCE AND SELF-EVALUATIVE BEHAVIOR: A CRITIQUE OF METHODOLOGY

DOUGLAS P. CROWNE AND MARK W. STEPHENS[1]
Ohio State University *Purdue University*

"Self-acceptance" has become a popular concept in psychological literature. Along with "rigidity," "authoritarianism," and "conformity," it has come to particular prominence in the last decade, perhaps reflecting an evolution in value systems in American culture. Concepts pertaining to the self have been given considerable space in the writings of personality theorists and social-personality psychologists and inevitably have found their way into psychological research.

Self-acceptance has been particularly identified with Rogers' personality theory and is accorded the status in that system of a major therapeutic goal. Phenomenological research on self-acceptance dates from the classic study of Raimy (1948). However, very similar concepts have played dominant roles in other theories—e.g., Snygg and Combs (1949), Horney (1950), and Sullivan (1953). More important, self-acceptance seems to have been pre-empted for less systematic, eclectic usage by a great many practicing clinicians and researchers (Cowen, 1956; Cowen, Heilizer, Axelrod, & Alexander, 1957; Zuckerman, Baer, & Monashkin, 1956; Zuckerman & Monashkin, 1957). The major portion of the research on self-acceptance derives from Rogers' self-theory, but studies based on other theories (Block & Thomas, 1955; Sarbin & Rosenberg, 1955) and the generally empirical investigations referred to above attest to the breadth of current interest in the behaviors subsumed under this broadly interpreted construct.

While no single definition of self-acceptance would be accepted by all who use the term, the phenomenological view of Rogers seems to represent at least a common point of departure. From the definition of a *self-concept* construct the concept of self-acceptance is derived, referring, at least operationally, to the extent to which this self-concept is congruent with the individual's description of his "ideal self."

The majority of self-acceptance tests have followed this model (see Table 1). A somewhat different psychometric model has been proposed by Gough (1955), in which self-acceptance is inferred from the ratio of "favorable" self-descriptive statements to the total number of self-descriptive statements made by the subject.

A common denominator in the definition of self-acceptance, judging from the operations employed in its assessment, would seem to be the degree of self-satisfaction in self-evaluation. This definitional consensus, however, is achieved at the level of operations, and other meanings may be implied by self-acceptance *constructs*. Phenomenological theorists, for example, appear to be interested in an "internal" phe-

[1] The authors would like to express their indebtedness to the following persons, who critically read this paper and made a number of valuable suggestions: Donald Campbell, Shephard Liverant, Julian Rotter, Lee Sechrest, Charles Smock, and Janet Taylor.

TABLE 1

Classification of Some Tests of Self-Acceptance

Name of Test	Type	Score Obtained
SIO (self-ideal-other) *Q* sort (Rogers & Dymond, 1954)	*Q* sort	Pearson correlation between sorts of self and ideal on 100 items. Also, "adjustment score" based on number of favorable statements placed on "like me" end of distribution and number of unfavorable statements placed on "unlike me" end.
Index of Adjustment and Values (Bills, 1958; Bills, Vance, & McLean, 1951)	Adjective rating scale	Self-acceptance score = sum of self-concept ratings (1–5 scale) on 49 traits. Also, a self-ideal discrepancy score is calculated. Norms available.
Adjective Check-List (Gough, 1955)	Adjective check list	Self-acceptance score = number of favorable adjectives checked divided by total number of adjectives checked.
Buss scale (Buss & Gerjuoy, 1957; Zuckerman & Monashkin, 1957)	Adjective check list	Sum of differences without regard to sign of scale values (based on psychologists' ratings) of adjectives checked on self and ideal descriptions.
Self-Rating Inventory (Brownfain, 1952)	Self-rating scale	"Positive self-concept" and "negative self-concept" scores. Self-acceptance = sum of positive self-concept description weights minus negative self-concept description weights, disregarding sign.
Attitudes toward Self and Others Questionnaire (Phillips, 1951)	Self-rating scale	Sum of weights (1–5) on each item. Norms available.
Berger Self-Acceptance scale (Berger, 1952)	Self-rating scale	Sum of item weights (1–5).
Interpersonal Check List (LaForge & Suczek, 1955)	Adjective check list	Intensity scale values for each adjective (1–4). Self-acceptance = discrepancy between self and ideal ratings.

nomenal state. Other theorists (Block & Thomas, 1955) have formulated self-acceptance as a function of an ego-control construct. The phenomenological concept of Rogers and the psychoanalytic set of meanings implied by Block and Thomas' construct of ego control probably diverge in important respects. The purpose here, however, is merely to illustrate the point that emphasis on definitional clarity achieved at an operational level tends to ignore the probably significant differences in the implied theoretical meanings of self-acceptance.

Reflecting in part the widespread interest in self-acceptance are the numerous instruments which have been devised to measure the construct. A striking phenomenon of research in this area is that these tests, characterized by a diversity of both theoretical and psychometric models, have apparently been assumed to be interchangeable. Thus, characteristic of self-acceptance research appears to be a basic conception that measures of this construct possess face validity: that is, in a simple denotative sense, the tests are viewed as being manifestly similar (Peak, 1953).

Criterion validation of self-acceptance tests is, of course, logically impossible, and attempts at construct validation do not lend much faith in the validity even of a particular test, much less of all the different tests. Face validity, however, has apparently been assumed without question. The acceptance of face validity—that is, manifest similarity—implies adherence to a further assumption incorporated in phenomenological theory—that of the validity of self-reports (Rogers, 1951, p. 494). In terms of these assumptions, a self-acceptance test is valid if it looks like a self-acceptance test and is similar to other tests, and what a person says about himself self-evaluatively is accepted as a valid indication of how he "really" feels about himself.

The acceptance of these assumptions, whether acknowledged or implicit, has definite implications for the assessment of self-acceptance and for the interpretation of experimental results in this area. This paper will show that there are four major problems in the measurement of this construct and that, in view of the common adherence to these assumptions, the results of studies on self-acceptance are rendered highly ambiguous. These issues seem, despite their essential pertinence to research on self-acceptance, to have been sufficiently ignored to warrant exposition in this paper. It will be seen that these issues are not limited solely to self-acceptance, but represent instead basic logical and psychometric considerations which may serve to illustrate problems in personality research in general.

Equivalence of Operations

As observed above, the diverse tests of self-acceptance have been assumed to be equivalent operations for measuring behaviors subsumed under the construct. The failure of experimenters to consider the problem of the equivalence of assessment operations in published reports (Bills, Vance, & McLean, 1951; Block & Thomas, 1955; Calvin & Holtzman, 1953; Cowen, Heilizer, & Axelrod, 1955; Hillson & Worchel, 1957; Phillips, 1951) raises the question of the basis on which the findings of individual studies employing different measuring operations are generalized and incorporated in the larger body of self-acceptance research. The basis of generalization, in view of the absence of explicit consideration of the question, must be inferred to lie in the assumption of face validity as defined above. Even statements implying differences among self-acceptance tests fail to deal with the logically sequential question of the extent to which these differences may mean that self-acceptance as measured by Test 1 is not the same as self-acceptance as measured by Test 2. The following excerpt illustrates this point (Cowen et al, 1955):

> Presumably each of these classes of [self-acceptance] measures has certain peculiar advantages and limitations. . . . In any case, a good many data have now been presented demonstrating some empirical validity for both types of measures (i.e., they can discriminate among subjects with respect to other personality and behavioral indices in a manner roughly consistent with predictive expectations based on phenomenological theory) (p. 242).

These writers do not make clear what relationship obtains between the classes of self-acceptance tests (tests yielding discrepancy scores versus self-concept rating devices) or, more basically, how phenomenological personality theory can lead to operations that apparently can satisfy certain predictions in the case of one class of instruments but requires different operations to obtain positive results from other hypotheses based on the same construct.

According to the notion of face validity, what looks like a test of self-acceptance *is* such, by definition. All the test constructor is required to do, in terms of this criterion, is to elicit self-evaluative statements from subjects. All measures that conform to this requirement achieve validity and are therefore equivalent. By this procedure the test itself becomes the construct, in the sense of the narrowest kind of operational definition.

An operational definition stating what is measured by a given device or procedure in terms of specified measurement operations is, of course, a perfectly legitimate and necessary procedure in scientific investigation *as long as the interpretation of results is strictly confined to the particular test or measurement procedure.* A problem arises, however, when an attempt is made to generalize from experimental findings with a particular test to results obtained by *different* assessment operations. The problem similarly occurs in another case when a certain test is applied to an experimental problem and negative results are interpreted as disconfirming the hypotheses relating the construct to observables. As Jessor and Hammond (1957) have pointed out, in the absence of an explicit, logical relationship between the superordinate construct and the operations designed to assess it, conclusions cannot be made concerning the validity of the hypotheses since invalid measurement operations could equally account for negative findings.

The point at issue is that tests of self-acceptance (or, for that matter, of any construct) which are based on different construct systems and in the development of which different procedures and items have been employed are not equivalent *in the absence of empirical demonstration of their relationships;* they must be shown to be either highly related to each other or similarly related to other constructs in the nomological net. Further, in the absence of demonstrated equivalence, experimental results cannot be generalized to findings with a different instrument. This seems to be so obvious a consideration that explication here is redundant. The fact remains, however, that the equivalence of self-acceptance tests has been assumed despite their independent derivation and despite the relative lack of empirical demonstration that there is a high degree of common variance among them.

In respect to the latter point, three studies are of interest. Bills (1958) reports a correlation of .24 between the self-concept score on the Index of Adjustment and Values (IAV) and the "self-score" of the Phillips Attitudes Toward Self and Others Questionnaire (1951). A correlation of .56 is reported between the Bills self-ideal discrepancy score and the Phillips self-score. Omwake (1954) found a correlation of .55 between the IAV self-acceptance (self-ideal discrepancy) score and the self-score on the Phillips questionnaire and a correlation of .49 between the self-acceptance score on the IAV and the Berger self-acceptance scale (Berger, 1952). In a recent study, Cowen (1956) found that two self-acceptance measures yielding self-ideal discrepancy scores (Bills IAV and the Brownfain Self-Rating Inventory) were uncorrelated. The magnitude of these correlations indicates that the prediction of scores on one of these measures from scores on another would be accompanied by a wide margin of error.

The diversity of item selection procedures, item content, type of response elicited, and test format which is characteristic of test construction in this area suggests that what is operationally defined as self-acceptance on one test may be quite different from the sample of self-evaluative behavior elicited in another psychometric situation. Further, self-acceptance is construed differently by different theorists (cf. Block & Thomas, 1955; Butler & Haigh, 1954; LaForge & Suczek, 1955; Sarbin & Rosenberg, 1955), and these definitional differences are undoubtedly reflected in self-acceptance tests.

Even if one grants the assumption of face validity with its clearly implied meaning of equivalence *as made by the experimenter*, to assume that subjects will perceive these psychometric situations in the same way is another matter. It is quite conceivable that subjects may categorize the self-evaluative situations represented by the various tests of self-acceptance quite differently, with the result that scores obtained on these measures will not be congruent. According to this argument, a subject's expectancies that his goals will be achieved or frustrated as a result of his sorting a number of statements on a forced-choice distribution from "like me" to "unlike me" (Butler & Haigh, 1954) may be quite different from the expectancies aroused by a situation in which he is asked to attribute certain adjectival characteristics to himself (Gough, 1955). Ironically, a phenomenological definition of a self-report variable is particularly obligated to account for differences in the subject's *perception* of the measurement device. In any case, unless it can be shown that there is a high degree of congruence of the various measures within the experimental populations sampled, one is without means of measuring self-acceptance as phenomenologically defined. The individual's private, unique experience of self-satisfaction or dissatisfaction remains, indeed, private.

It seems highly probable that differences among self-acceptance tests plus the likelihood that subjects will categorize these tests differently may result in the sampling of relatively nonoverlapping behaviors by the various tests. To be recognized is the fact that this is an empirical problem for which, to the writers' knowledge, the three studies cited above provide the only suggestive evidence.[2] The recently proposed model (Campbell & Fiske, 1959) for assessing convergent and discriminant validity would seem to be highly appropriate for determining the tenability of the assumption of equivalence of operations for measuring self-acceptance.

Definition of the Construct

Specifying Parameters

The ability to reach generalized conclusions from current self-acceptance research seems to be limited by a failure to give adequate definitions to the construct itself. As Rotter

[2] Since the completion of this article, further research has been published which bears directly on the problem of the equivalence of self-acceptance tests and suggests that a socially desirable response set may constitute a major source of variance (Crowne, Stephens, & Kelly, 1961).

(1954) has pointed out, it is important to distinguish between ideal, theoretical, and operational definitions of a given construct. An experimenter can define self-acceptance, for example, as the behavior sample (or as the "internal" phenomenal state *reflected* by the behavior sample) obtained on a particular test. But he is usually not interested in restricting his interpretation of his findings (if any) to this limited behavior sample, and he seeks to place his results in the larger context of research by other investigators and to generalize his findings to "real life" situations such as those encountered in clinical practice. By a narrow interpretation of operationism, the experimenter has made it logically indefensible to relate his findings to a theoretical system, to results obtained with other measurement devices, or to "real life" situations. When nothing more than an operational definition is offered, the parameters defining the variable are not specifiable, and there is no basis for generalization of the results.

At the other extreme, definitions of self-acceptance at an abstract level, not specifically articulated with other variables in a theory or tied to a specific test, are apt to be semantically loose and to be subject to differing interpretations. It is true, of course, that definitions of variables at this level transcend any particular set of operations and can usually be applied to an infinite variety of situations and behaviors. The looseness of such definitions, however, precludes rigorous tests of hypotheses and makes precise communication impossible. In self-acceptance research there have been few if any definitions of the construct which are not either rigidly operational or highly abstract.

The deduction from an abstract definition, with all its surplus meanings, to specific operations is likely to be a tenuous one and, perhaps more often than not, is a private, nonrepeatable process. An intervening step is necessary in which the construct is broadly defined in terms of specific behavioral referents and preferably in relation to other variables in a specific theory. A "working definition," as Rotter has defined it, clearly represents an attempt to specify the parameters of the variable in question so that both generality and precise communication are gained. Self-acceptance research appears to have lacked such definitions.

Although this paper is chiefly concerned with pointing out certain methodological pitfalls in research on self-acceptance, some clarification may be achieved by defining briefly this intermediate theoretical step and attempting to relate the logic of construct validation to the more general theoretical problem. Rotter's working definition could be described as a *definition at the construct level.* In terms of this view, the behavioral referents *and the hypothesized relationships* of the construct are described as part of its definition—that is, the implied meanings of the term are publicly specified. In effect, specifying the behavioral referents and hypothesized relationships reduce to the same thing: locating the construct in a nomological net. In the language of test construction, Cronbach and Meehl (1955) write:

> Construct validation takes place when an investigator believes that his instrument reflects a particular construct to which are attached certain meanings. The proposed interpretation generates specific testable hypotheses, which are a means of confirming or disconfirming the claim. . . . To validate a claim that a test measures a construct *a nomological net surrounding the concept must exist* [italics added] (pp. 290–291).

The logic of construct validation cannot be invoked to justify the identification of a particular set of operations as unique to a given construct, nor does it support the view that a construct is "validated" by the confirmation of a single hypothesis. The establishment of a single relationship belongs more properly in the domain of criterion oriented validity, as Cronbach and Meehl point out. With construct validation procedures clearly at issue, it would seem to be desirable to specify in advance the referents of self-acceptance. When the situations in which the behaviors subsumed under the construct and the behaviors themselves are identified, some idea of the generality and functional unity of self-acceptance is afforded, and relationships to other constructs, situations, and measurement operations can be suggested at a logical level.

Underwood (1957) has described the difficulty in moving from theoretical definitions (or constructs) to operational definitions—a difficulty that appears to be characteristic of psychological research. Campbell and Fiske (1959) have extended Underwood's point to show that the transition from operations to construct can involve perplexities equally difficult. The essence of the latter problem is that a single set of operations is capable of multiple interpretations; convergence on a single interpretation (that is, establishing that a relationship holds in a particular nomological net and cannot be more adequately accounted for in another net) is achieved by a process of triangulation from a number of different operations. Convergent validation, however, involves complex designs and extensive preliminary research efforts. Further, convergent validation does not necessarily help to make more explicit the descent from a theoretical model to measurement operations. According to the present view of definition at the construct level, this explicitness would be achieved and the reverse problem, that of interpreting results from a set of operations, might be at least partially solved. That is, alternative explanations of experimental findings could be examined in the light of the hypothesized relationships proposed in different construct systems claiming to explain the same body of data, with the result that incomplete or inconsistent interpretations might be discarded in favor of interpretations whose "fit" to the data is more adequate.

For example, phenomenological theory implicitly hypothesizes a linear relationship between self-acceptance and adjustment (Butler & Haigh, 1954), while acknowledging the possibility that very high reported self-acceptance may indicate "defensive" unwillingness to reveal personal dissatisfaction. Block and Thomas (1955), however, have shown that a curvilinear model, in which both very high and very low self-acceptance are associated with maladjustment, affords a better explanation of the phenomenon of defensiveness. It is conceivable that more explicit formulation of the phenomenological self-acceptance construct and its derived test procedures might have provided a more adequate explanation of defensive responding in the Butler and Haigh study. More precise definition of the variable in question might thus have directed a search for operations less susceptible to systematic response bias.

In a recent paper, Cowen and Tongas (1959) have reviewed a number of construct validation studies on the IAV (Bills, 1958). They point to the fact that several of these studies have reported significant results in the direction opposite to theoretical expectation. In one study, on 10 of 21 hypotheses specifying differences between high and low self-acceptance scorers, many differences were found which indicated that subjects with high self-acceptance scores were more maladjusted than low scorers (Bills, 1953a). As Cowen and Tongas observe, high self-acceptance should theoretically be associated with satisfactory adjustment, not maladjustment. Another theoretical inconsistency occurred in the failure to show that lowered self-concept ratings and longer response times in word association are associated with conflict and emotionality. The results of this study were again, in fact, significant in the opposite direction (Bills, 1953b). Bills interpreted these findings as indicating a decrease in defensiveness. Cowen and Tongas argue, however, that:

> Unless procedures can be specified before the fact, by which we can discriminate the high SC (self-concept) score representing good adjustment from the high SC score representing defensiveness, we are operating within a closed system in which the results of a given experiment, irrespective of their direction, can be interpreted as confirming the underlying theory (pp. 362–363).

Self-acceptance research is in need of clear construct-level definitions in which the relationships of the construct to other variables are explicitly stated. These definitions must refer primarily to the relationship of self-acceptance to other variables in the general theory in which the construct is embedded. Depending upon the particular theory, definitions might specify the nature of the relationship of self-acceptance to adjustment; to such personality variables as creativity, neuroticism, and defensiveness; to interpersonal variables such as acceptance of others; to environmental, social, and cultural variables, as, for example, the role of cultural sanctions in self-evaluation, or the influence of the experimental (or therapeutic) context on self-appraisal.

Representative Sampling

A second problem associated with the definition of the parameters of self-acceptance concerns the representative sampling of self-acceptance test items. As applied to the construct of self-acceptance, the problem of representative sampling is involved in the systematic sampling of some specified universe of self-evaluative behaviors. Assuming that one has defined this population theoretically, it is then of importance to draw one's sample of test items in such a way as to represent their occurrence in the population. The achievement of representative sampling in this respect means that generalization can reasonably be attempted to other situations and/or behaviors than those of a particular experiment or test. Although the behavioral referents of self-acceptance might seem obvious, on closer scrutiny it appears that there is notable confusion resulting from a lack of consensus as to what these referents are.

Some examples from published research may illustrate what is implied by failure to sample representatively a population of self-evaluative behaviors. Butler and Haigh (1954) begin with Rogers' abstract definition of the self-concept. Then, they write:

> A set of one hundred [self-reference] statements was taken at random from available therapeutic protocols. (Actually, the statements were selected on the basis of accidental, rather than random, sampling) (p. 57).

The population of relevant self-percepts was therefore restricted to those verbalized by some sample of clients in client centered therapy, the basis for sampling was accidental, and thus there is no precise definition of self-acceptance in terms of what particular self-percepts define its parameters. The finding that changes in self-acceptance were demonstrated to occur as a function of client centered therapy is thereby limited to the particular conditions of this experiment, the subject population used, and the particular items employed in the *Q* sort measure. For example, it is quite possible (but unknown) that the statements used comprise a sample biased in favor of client centered counseling as perceived and defined by the judges (presumably Butler & Haigh) who selected the items.

A second example can be seen in the development of the IAV (Bills et al., 1951). The items (adjectives) in the IAV were drawn from Allport and Odbert's (1936) list of 17,953 traits. The basis of selection was the frequent appearance of the adjective in question in client centered interviews and whether it presented a "clear example of self-concept definition." Self-evaluation on the IAV, then, pertains only to the Allport and Odbert traits mentioned frequently in client centered interviews, and generalization to other self-evaluative situations, or traits, would be tenuous.

Gough's (1955) Adjective Check List (ACL) affords a third illustrative example. The ACL consists of 300 adjectives selected from Cattell's (1943, 1946) consolidation and factorization of the Allport and Odbert trait list. The basis on which the 300 adjectives in the ACL were derived from Cattell's list of 171 trait variables is not specified. In addition, the assumptions both of Allport and Odbert in their original derivation of the trait list and of Cattell in his factorization are further restrictions in interpreting ACL scores.

With such lists of traits or items it is necessary to assume either that they truly represent *all* self-percepts, or at least that they represent the most important ones. But, especially for the phenomenologist, must it not be assumed that these are different for different subjects and/or subject populations? Must not this list, then, be tailor-made to the subject to be truly representative for him (a totally idiographic procedure)? Perhaps what is required is that the subject generate his own list of self-descriptions, or a self-description, and the values he attaches to the separate elements and to the composite. Kelly's (1955) Role-Construct Repertory Test appears to fit this model.

It would seem possible to achieve some degree of representativeness in the sampling of a defined universe of self-reference items. The definition of the population is properly referable to the theory in which the self-acceptance construct is embedded. That is, one should be able to deduce from the theory the nature of the items to be sampled (although, from a phenomenological theory, one might protest that this population of items is unique to the individual; but this only thickens the soup). Not only should the population of subjects be specifiable (for example, the theory has particular relevance to persons in client centered therapy), but what constitutes a relevant self-evaluative statement (that is, the basis for self-evaluation) should be deducible as well. The relative adequacy of theories employing self-acceptance constructs is clearly at issue in this case.

With regard to the problem of sampling a defined universe, one approach has been suggested by Crowne (1959). Definitions of self-accepting and self-derogatory behavior from the point of view of social learning theory (Rotter, 1954) were given first to psychologist judges and then to judges drawn from the subject population (introductory psychology students) to which generalization was intended. The psychologists were asked to generate from these definitions lists of self-evaluative behaviors—that is, behavioral referents, or cues, of self-acceptance and self-rejection—common in such a subject population. Subject judges were given a list of 300 adjectives (actually, the ACL) and asked to rate each adjective in terms of the extent to which they felt that, if it were checked by one of their peers as descriptive of himself, self-acceptance or self-rejection would be indicated. Items were then selected on the basis of high interjudge agreement of both psychologist and subject judges. In this way the items were tied to, and representative of, both the superordinate theory and the specific population of self-evaluative behaviors common to the experimental population. This procedure was still limited, however, to the extent that the list of 300 adjectives failed to represent some clearly defined universe. Generalizing the procedures used in this study, it would be possible to elicit descriptions of self-acceptance and self-rejection (the definitions for the judges being derived from theory) from a large sample of judges drawn from the appropriate population. Items might then be selected from descriptive units on which there was high interjudge agreement. The methodological and psychometric considerations proposed by Jessor and Hammond (1957) would presumably dictate the form of the scale, type of response, and related aspects of test construction.

This section has been concerned with two problems related to the definition of the parameters of self-acceptance: (*a*) the necessity of providing a definition at the construct level, in which the behavioral referents of self-acceptance are specified and the construct located in a nomological net; and (*b*) the need to consider the representativeness of the sampling of a population, as defined in *a* of self-reference statements or items. Failure to meet these criteria results in the inability of the experimenter or test constructor legitimately to generalize from the particular conditions (subjects and stimuli) of his experiment or test.

Social Desirability

The third general issue to raise concerns the extent to which self-evaluative responses are influenced by "defensive behavior" (Butler & Haigh, 1954; Zuckerman & Monashkin, 1957), "self-protective response tendencies" (Crowne, 1959), or "social desirability" (Edwards, 1957; Kenny, 1956). It is important, however, first to consider whether these terms refer to the same or different phenomena.

Butler and Haigh apply the term "defensive responding" to the responses of those individuals who do not reveal the extent of their self-dissatisfaction and who, by other criteria, would be judged as maladjusted. (These authors thus seem to reject, for some subjects at least, the assumption of validity of self-reports, although how this can be done within a phenomenological frame of reference is hard to understand.) "Defensiveness" has been used by Zuckerman and Monashkin to refer to the phenomenon whereby "The person who is self-satisfied is likely to answer MMPI items in a way which he considers personally and socially desirable" (p. 147). Crowne used the term "self-protective behavior" to refer to the unwillingness of some individuals to acknowledge self-dissatisfaction. These three terms, then, have been used to refer to highly similar phenomena.

"Social desirability" as defined by Edwards (1957), however, refers primarily to the

> scale value for any personality statement such that the scale value indicates the position of the statement on the social desirability continuum (p. 3).

It also applies, as Edwards further points out,

> to the tendency of all subjects to attribute to themselves, in self-description, personality statements with socially desirable scale values and to reject those with socially undesirable scale values (p. vi).

Whereas the above concepts of "defensiveness" have been applied to the motivation, presumably greater for some subjects than for others, to conceal self-dissatisfaction, Edwards' notion of "social desirability" refers to a characteristic of *items*—that is, their location on a continuum of social desirability, which determines the proportion of subjects who will attribute the characteristics to themselves.

Butler and Haigh, and also Zuckermen and Monashkin, conclude that subjects who are unwilling to attribute undesirable characteristics to themselves or confess self-dissatisfaction are by that very fact maladjusted, and presumably therefore self-dissatisfied. This, however, is obviously an hypothesis for investigation, and not necessarily true by definition. Self-acceptance tests do not directly indicate whether the subject is *willing* to express self-discontent, but only whether he *does* express it. Zuckerman and Monashkin have also suggested, in fact, that subjects giving more socially undesirable responses may have a different conception of what *is* socially desirable, and thus they implicitly suggest that these subjects may actually *not* differ in terms of their *need* to respond in a socially desirable fashion. Such a difference in conception of what is socially desirable might be expected to be associated with maladjustment, but it would certainly be a less direct indication of self-dissatisfaction per se.

Four separate hypotheses could be advanced concerning the relationship between social desirability and responses on self-acceptance (or any other self-report) tests. Each of these is capable in some degree of being tested.

Hypothesis I. Social desirability has no effect on test responses. This is essentially the assumption of validity of self-reports: that what the subject says about himself is a valid and direct indication of what he feels or thinks, at least at the time, about himself. This, incidentally, seems to be a necessary *assumption* for phenomenologists, although it is a testable proposition.

Hypothesis II. Social desirability factors account for equal variance in all subjects' test scores. This assumption is tenable from Edwards' approach and could be held even in the face of most of the research to be reported below. It posits, in effect, that once one has accounted for variance due to nomothetically determined social desirability in any subject's test score, what is *left* indicates the subject's true self-feelings.

Hypothesis III. Social desirability, while it may or may not be an important factor for all subjects, accounts for *more* of the variance for some subjects than for others. This corresponds to the suggested differences in *need* to perform in a socially desirable way, protect the self, and disguise self-discontent. It is interesting that such need has been supposed to be an important variable only for those who show relatively high self-acceptance or social desirability scores: the rebel, or the individual seeking succorance, may produce very *low* scores, as a result of a complementary need to perform in a socially *undesirable* way, and still not necessarily differ from others in terms of over-all adjustment *or* "true" self-acceptance. In any case, such a conception as this suggests research determining the correlates of this need to perform in a socially desirable, or to perform in a socially undesirable, way.

Hypothesis IV. Variance associated with a nomothetically determined social desirability factor reflects differences in the conception of what *is* socially desirable. This hypothesis is not necessarily in conflict with Hypothesis III: both factors could operate simultaneously, although separating the variance due to each might be quite difficult. This, as well as Hypothesis III, is definitely incompatible with Hypotheses I and II.

With the above distinctions in mind, then, the results of some investigations of the relationship of the social desirability variable to self-acceptance test scores can be examined. Kenny (1956) gave 25 self-descriptions previously employed in a study by Zimmer (1954) to a group of judges for social desirability scaling. Three independent samples of subjects then responded to these items in the form of a questionnaire, a self-descriptive rating scale, and a *Q* sort. The correlations between the social desirability scale values and the scores obtained on the questionnaire, rating scale, and *Q* sort were .82, .81, and .66, respectively. The last two correlation coefficients are based on a "real self" scores. Social desirability correlated .82 with the "ideal self" rating scale score and .59 with the "ideal" self *Q* sort.

Edwards (1955, 1957) and Edwards and Horst (1953) have also shown that *Q* sorts are highly influenced by the social desirability variable. In a study reported in 1955 and reviewed in 1957, Edwards found correlations of .84 and .87 for males and females, respectively, between item placement on a *Q* sort and the social desirability scale values of the items. In this case, the items were those employed in the development of the Edwards Personal Preference Schedule (1953).

In one study (Kogan, Quinn, Ax, & Ripley, 1957), a social desirability scale value-response correlation of .67 was found in a hospitalized psychiatric patient sample diagnosed as psychoneurotic. The correlation in a control group of male college students was .85. It is interesting to speculate upon the possible significance of the difference in the magnitude of the correlation between self-description and social desirability values found for the patient and nonpatient groups. Perhaps, as Hypothesis IV proposes, maladjusted persons have different conceptions of social desirability in self-evaluative situations.

Studies by Berger (1955), Block and Thomas (1955), and Zuckerman and Monashkin (1957) are also relevant to the problem of social desirability. These studies investigated the relationships between self-acceptance and the clinical and "validity" scales of the MMPI. Employing different subject populations—college undergraduate students in the first two studies and hospitalized psychiatric patients in the latter investigation—and different measures of self-acceptance, there was nevertheless considerable agreement in the findings. Self-acceptance was found to be significantly negatively correlated with a number of the clinical "adjustment" scales and positively correlated (*r*'s ranging from .33 to .58) with the *K* scale, interpreted as a measure of test-taking defensiveness (McKinley, Hathaway, & Meehl, 1948). Zuckerman and Monashkin took their findings to mean that "both self-acceptance and MMPI scales are probably being influenced more by the common trait of defensiveness than by actual adjustment" (p. 147). The term "defensiveness," with its connotation of maladjustment, seems less applicable here than "social desirability," especially in view of the high correlation (.81) reported by Edwards (1957) between the *K* scale and his Social Desirability Scale. With approximately 65% of the variance accounted for in the covariation of these two scales, the results of the three studies cited above would seem to be a function of the common denominator of social desirability. Thus, the items on the self-acceptance tests used and those on the MMPI are highly related to the scale values on Edwards' Social Desirability Scale.

In the study referred to earlier, Cowen and Tongas found a correlation of .91 between social desirability ratings and the self-concept score of the IAV. A correlation of .96 was obtained between social desirability ratings and the ideal-self score on the IAV. The latter correlation might be taken to suggest culturally stereotyped conceptions of what one ought to be that would be consistent with Hypothesis IV above. In another investigation (Nebergall, Angelino, & Young, 1959), it was found that subjects who reported high self-acceptance tended to disagree with group judgments of adjustment. For most subjects, in fact, self-acceptance ratings were higher than group ratings. Again, these findings may be understood in terms of the individual's need to present himself in what he regards as a culturally sanctioned manner.

While this discussion has been concerned primarily with the social desirability factor in self-acceptance tests, it seems highly probable that any self-report device will be affected by the social desirability of items or of available responses. Failure to control for the effects of this variable by one of several available procedures (Edwards, 1957) means, in effect, that the test in question may better be interpreted as a measure of social desirability (that is, the subject's conception of social desirability or need to perform according to it) than of self-acceptance. This can be illustrated by means of an hypothetical experiment. It might be hypothesized that need-determined perceptual behavior—for example, perceptual reactivity to threat—is related to self-acceptance (cf. Cowen et al., 1957). Failure to control for social desirability in the self-acceptance assessment operations would make the results, no matter what the outcome, uninterpretable in terms of self-acceptance. In the light of what is already known about the influence of social desirability on self-report devices, the most probable interpretation of such an experiment would be that perceptual reactivity to threat is related (or unrelated) to the socially desirable responding of subjects—that is, their need to be perceived in a particular way or their conception of how they want to be perceived. Not provided in this experiment are the operations for determining the relationship between perceptual reactivity and "real" self-acceptance.

While studies of the effect of the social desirability variable on many of the commonly employed tests of self-acceptance have not been done, the results of the investigations discussed above would suggest that self-evaluative tests are particularly susceptible to criticism on social desirability grounds. A common denominator in research findings on self-acceptance may well be the variable of social desirability. Edwards (1957) and Jackson and Bloomberg (1958) have made a similar analysis with respect to the Taylor anxiety scale (Taylor, 1953). Systematic investigation of both the parameters and the effects on test behavior of social desirability would clearly seem to be in order. That self-acceptance tests are influenced by factors other than the manifest content of the items, however, seems beyond dispute.

THE GENERALITY OF SELF-ACCEPTANCE

To this point the issues discussed have been pertinent strictly to psychometric and methodological problems in assessing self-acceptance. A further issue to be raised, although it certainly has methodological ramifications, is the primarily theoretical question of the generality of self-acceptance.

Generality involves two related problems, one empirical and the other a theoretical problem of interpretation. Empirically, there is need of evidence concerning the temporal stability of self-acceptance; the consistency of an individual's self-acceptance from one situation to another (for example, in friendly vs. hostile groups, or where self-effacement is rewarded or not rewarded); the generality of self-acceptance in reference to different aspects of the "self" (for example, in reference to morality vs. in reference to interpersonal effectiveness); and agreement of different kinds of *manifestation* of self-acceptance (for example, spontaneous self-appraisal vs. that manifested in an undisguised test such as the ACL vs. inferences drawn from a TAT protocol). The theoretical question is simply how best to construe self-acceptance. If, as has been suggested (Rogers, 1951), the self-concept and self-acceptance can be considered to be relatively stable characteristics of a person, one should find that situational variables have only a negligible effect on self-acceptance, that measures of self-acceptance taken in different social contexts are highly correlated, and that measures taken over temporal intervals are likewise highly stable. If these questions can be answered positively, it would be reasonable to construe the self-concept, from which the discrepancy notion of self-acceptance is derived, as a meaningful variable on which there are consistent differences between subjects, and it would be highly appropriate to think of individuals in terms of their characteristic levels of self-acceptance. To the degree that self-acceptance is a function of variables associated with specific situations or types of situations, however, it will be more fruitful to investigate self-evaluative behavior per se and its situational determinants.

The empirical evidence with respect to the generality of self-acceptance is rather scanty. The fact that studies have not attacked this question may be attributable to the general assumption that self-acceptance *is* consistent. Three investigations have been reported which do bear on this question. With respect to temporal stability, Taylor (1955) reports a test-retest correlation of .79 (presumably based on self-sort–ideal-sort discrepancy scores) over an interval of a week. Butler and Haigh (1954) report the correlations between self-sorts and ideal-sorts for each subject in a control group ($N=16$) not receiving therapy for two Q sort administrations separated by a considerable period of time. Although consistency was apparent, Butler and Haigh noted that

> there are some sharp individual changes which indicate that alteration in self-ideal congruence does occur at times in the absence of therapy (p. 67).

Concerning the influence of situational variables of self-acceptance, a study by Thorne (1954) is relevant. Employing the IAV, Thorne found that following induced failure on a mirror drawing task, subjects whose initial level of self-acceptance was high tended to lower their self-ratings in the direction of a more realistic evaluation, while originally low self-accepting subjects tended to increase self-acceptance scores and showed concern over loss of self-esteem. The results of this study would suggest that self-acceptance is influenced by environmental events and that persons respond self-reflexively to perceived successes and failures.

It would appear, from this brief discussion, that studies should be devoted to the problem of the generality of self-evaluative behavior. Of particular interest are the questions of temporal stability, influence of situational variables, and the effect on self-evaluation of such factors as success, failure, and punishment.

Self-Acceptance vs. Self-Evaluative Behavior

It has been necessary at several points in this discussion to point out differences between a phenomenological and a behavioristic approach to self-acceptance. Since these differences are basic to the research approaches—not to mention the way in which such research is construed—in this general area of inquiry, and since these differences seem not to have been fully appreciated by all who have written on the topic, some further discussion of them is in order.

A phenomenological approach to self-acceptance is concerned with self-acceptance itself, or "real" self-acceptance, as a totally private, subjective experience of the subject. By definition this is never observable by any other; the best that an experimenter or clinician can hope to do is make relatively accurate guesses, or inferences, concerning the existence, or degree, of the variable as it "exists" in the subject. By such a definition, self-acceptance corresponds to MacCorquodale and Meehl's (1948) early conception of an "hypothetical construct"—something which cannot be observed but still is assumed to exist—except that there is little suggestion that self-acceptance even *can* be observed by anyone other than the subject himself. It is only with some difficulty, it would seem, that a phenomenologist can avoid the necessity of assuming the validity of self-reports. Representative sampling, and also an idiographic procedure for determining what are the most salient aspects of a subject's self-evaluation, would seem to be most important in a phenomenological approach to the assessment of self-acceptance. Social desirability, on the other hand, should be assumed *not* to be a factor in self-reports. To assume a high degree of generality or consistency—temporal, situational, etc.—is not necessarily essential to a phenomenological approach; however, in any theory which posits generalized self-acceptance as an important dimension on which to compare people, empirically determined generality of the variable is, naturally, crucial.

A behavioristic concern with self-acceptance might more clearly be directed toward "self-evaluative behavior," on the other hand. The additional inference of some underlying, real if unobservable, phenomenological state is not essential to a study of self-evaluative behavior per se; and it might be pointed out that self-evaluative behavior is an interesting and perhaps important focus of interest in and of itself. In such an approach, the assumption of validity of self-reports is clearly not essential; a clear construct-level definition of self-evaluative behavior, on the other hand, is. Generality, representative sampling in test construction, and the related question of equivalence of assessment operations are crucial questions only if the goal is to approach self-evaluative behavior as a trait, or consistent behavioral tendency, by which to classify people in a generalized fashion. It is quite feasible to examine self-evaluative behavior as a situationally determined phenomenon, or as one determined by a situation-person interaction, rather than as a trait. Social desirability, defensiveness, etc., become merely other variables related (or unrelated) to self-evaluative behavior, and not components of error variance. And, most important, it becomes an empirical matter to determine correlates (such as "adjustment") of various forms of self-evaluative behavior, either in general or in specified contexts.

This discussion is not meant to imply that a phenomenological interest in self-acceptance is unsophisticated or unworthy. Theoretical understanding of phenomenal states is a problem of inference. A clearer conception of "internal" phenomenal states such as self-acceptance would seem to be best derived from the observable behaviors of the person—that is, his self-evaluative behaviors. Phenomenological research would appear, in fact, to involve complexities that do not attach to more behavioristic efforts.

Summary and Conclusions

"Self-acceptance" promises to become an increasingly attractive focus of interest in both formal and informal psychological theory. A considerable volume of research has already been devoted to the topic, and a sizeable number of tests devised for such research. To this date, however, research has contributed an unknown, but perhaps very small, amount of understanding of self-acceptance and its relationships to other personality variables. The failures of self-acceptance research can be traced, at least in large part, to neglect of several crucial psychometric and methodological principles: the unsupported assumption of equivalence of assessment procedures, the absence of any clear construct-level definition of the variable, failure to construct tests in accord with principles of representative sampling, and questions concerning the social desirability factor in self-report tests. In addition, the absence of data concerning the generality of self-acceptance makes research results even more difficult to interpret; and the implications of the difference between a phenomenological approach to self-acceptance and a behavioristic approach to "self-evaluative behavior" have not been clearly understood.

The relative absence of systematic efforts in test development, standardization, and validation in this area is perhaps due to the fact that the focus of self-acceptance research to date has been chiefly on the preliminary testing of hypotheses, rather than the development of adequate tests as a primary aim. A test designed solely for the purpose of testing one or two hypotheses does not, it might be argued, require so much care as a test designed to serve as a standardized instrument for many purposes. Indeed, such an argument would continue, this care and time are not usually appropriate for such restricted purposes. (The development, use, and subsequent misuse of the Taylor Manifest Anxiety Scale would serve as a case in point as Taylor herself, 1956, has pointed out.) But when such tests are then used in further research as if they had been carefully and adequately constructed, little can ensue but error and confusion. And such seems to be the case in self-acceptance research.

Perhaps it is true that these tests are not yet used commonly in clinical settings where their inadequacies could lead to disservice to the client; perhaps it is true that the tests are used for very little other than research. But this only makes rigorous test construction the more important if research in such a complex area is to produce dependable and unambiguous results.

REFERENCES

Allport, G. W., & Odbert, H. S. Trait names: A psycholexical study. *Psychol. Monogr.*, 1936, **62** (1, Whole No. 211).

Berger, E. M. The relation between expressed acceptance of self and expressed acceptance of others. *J. abnorm. soc. Psychol.*, 1952, **47**, 778–782.

Berger, E. M. Relationships among acceptance of self, acceptance of others and MMPI scores. *J. counsel. Psychol.*, 1955, **2**, 279–284.

Bills, R. E. Rorschach characteristics of persons scoring high and low in acceptance of self. *J. consult. Psychol.*, 1953, **17**, 36–38. (a)

Bills, R. E. A validation of changes in scores on the index of adjustment and values as measures of changes in emotionality. *J. consult. Psychol.*, 1953, **17**, 135–138. (b)

Bills, R. E. *Manual for the Index of Adjustment and Values. Form: Adult and high school senior.* Auburn: Alabama Polytechnic Institute, 1958.

Bills, R. E., Vance, E. L., & McLean, O. S. An Index of Adjustment and Values. *J. consult. Psychol.*, 1951, **15**, 257–261.

Block, J., & Thomas, H. Is satisfaction with self a measure of adjustment? *J. abnorm. soc. Psychol.*, 1955, **51**, 254–259.

Brownfain, J. J. Stability of the self-concept as a dimension of personality. *J. abnorm. soc. Psychol.*, 1952, **47**, 597–606.

Buss, A. H., & Gerjuoy, H. The scaling of terms used to describe personality. *J. consult. Psychol.*, 1957, **21**, 361–369.

Butler, J. M., & Haigh, G. V. Changes in the relation between self-concepts and ideal-concepts. In C. R. Rogers & Rosalind F. Dymond (Eds.), *Psychotherapy and personality change.* Chicago: Univer. Chicago Press, 1954.

Calvin, A. D., & Holtzman, W. H. Adjustment and discrepancy between self concept and inferred self. *J. consult. Psychol.*, 1953, **17**, 39–44.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.*, 1959, **56**, 81–105.

Cattell, R. B. The description of personality: II. Basic traits resolved into clusters. *J. abnorm. soc. Psychol.*, 1943, **38**, 476–507.

Cattell, R. B. *Description and measurement of personality.* Yonkers-on-Hudson, New York: World Book, 1946.

Cowen, E. L. An investigation of the relationship between two measures of self-regarding attitudes. *J. clin. Psychol.*, 1956, **12**, 156–160.

Cowen, E. L., Heilizer, F., & Axelrod, H. S. Self-concept conflict indicators and learning. *J. abnorm. soc. Psychol.*, 1955, **51**, 242–245.

Cowen, E. L., Heilizer, F., Axelrod, H. S. & Alexander, S. The correlates of manifest anxiety in perceptual reactivity, rigidity, and self concept. *J. consult. Psychol.*, 1957, **21**, 405–411.

Cowen, E. L., & Tongas, P. N. The social desirability of trait descriptive terms: Applications to a self-concept inventory. *J. consult. Psychol.*, 1959, **23**, 361–365.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. *Psychol. Bull.*, 1955, **52**, 281–302.

Crowne, D. P. The relation of self-acceptance behavior to the social learning theory construct of need value. Unpublished doctoral dissertation, Purdue University, 1959.

Crowne, D. P., Stephens, M. W., & Kelly, R. The validity and equivalence of tests of self-acceptance. *J. Psychol.*, 1961, **51**, 101–112.

Edwards, A. L. *Manual for the Edwards Personal Preference Schedule.* New York: Psychological Corporation, 1953.

Edwards, A. L.. Social desirability and *Q* sorts. *J. consult. Psychol.*, 1955, **19**, 462.

Edwards, A. L. *The social desirability variable in personality assessment and research.* New York: Dryden, 1957.

Edwards, A. L., & Horst, P. Social desirability as a variable in *Q* technique studies. *Educ. psychol. Measmt.*, 1953, **13**, 620–625.

Gough, H. G. *Reference handbook for the Gough Adjective Check-List.* Berkeley: Univer. California Institute of Personality Assessment and Research, 1955. (Mimeo)

Hillson, J. W., & Worchel, P. Self concept and defensive behavior in the maladjusted. *J. consult. Psychol.*, 1957, **21**, 83–88.

Horney, Karen. *Neurosis and human growth.* New York: Norton, 1950.

Jackson, D. N., & Bloomberg, R. Anxiety: Unitas or multiplex? *J. consult. Psychol.*, 1958, **22**, 225–227.

Jessor, R., & Hammond, K. R. Construct validity and the Taylor anxiety scale. *Psychol. Bull.*, 1957, **54**, 161–170.

Kelly, G. A. *The psychology of personal constructs.* New York: Norton, 1955.

Kenny, D. T. The influence of social desirability on discrepancy measures between real self and ideal self. *J. consult. Psychol.*, 1956, **20**, 315–318.

Kogan, W. S., Quinn, R. D., Ax, A. F., & Ripley, H. S. Some methodological problems in the quantification of clinical assessment by *Q* array. *J. consult. Psychol.*, 1957, **21**, 57–62.

LaForge, R., & Suczek, R. F. The interpersonal dimension of personality: III. An interpersonal check list. *J. Pers.*, 1955, **24**, 94–112.

MacCorquodale, K., & Meehl, P. E. On a distinction between hypothetical constructs and intervening variables. *Psychol. Rev.*, 1948, **55**, 95–107.

McKinley, J. C., Hathaway, S. R., & Meehl, P. E. The MMPI: VI. The *K* scale. *J. consult. Psychol.*, 1948, **12**, 20–31.

Nebergall, Nelda S., Angelino, H., & Young, H. H. A validation study of the self-activity inventory as a predictor of adjustment. *J. consult. Psychol.*, 1959, **23**, 21–24.

Omwake, Katherine T. The relation between acceptance of self and acceptance of others shown by three personality inventories. *J. consult. Psychol.*, 1954, **18**, 443–446.

Peak, Helen. Problems of objective observation. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences.* New York: Dryden, 1953.

Phillips, E. L. Attitudes towards self and others: A brief questionnaire report. *J. consult. Psychol.*, 1951, **15**, 79–81.

Raimy, V. C. Self reference in counseling interviews. *J. consult. Psychol.*, 1948, **12**, 153–163.

Rogers, C. R. *Client-centered therapy.* Boston: Houghton-Mifflin, 1951.

Rogers, C. R., & Dymond, Rosalind F. (Eds.) *Psychotherapy and personality change.* Chicago: Univer. Chicago Press, 1954.

Rotter, J. B. *Social learning and clinical psychology.* New York: Prentice-Hall, 1954.

Sarbin, T. R., & Rosenberg, B. G. Contributions to role-taking theory. *J. soc. Psychol.*, 1955, **42**, 71–81.

Snygg, D. S., & Combs, A. W. *Individual behavior: A new frame of reference for psychology.* New York: Harper, 1949.

Sullivan, H. S. *The interpersonal theory of psychiatry.* New York: Norton, 1953.

Taylor, D. M. Changes in the self concept without psychotherapy. *J. consult. Psychol.*, 1955, **19**, 205–209.

Taylor, Janet A. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, **48**, 285–290.

Taylor, Janet A. Drive theory and manifest anxiety. *Psychol. Bull.*, 1956, **53**, 303–320.

Thorne, R. B. The effects of experimentally induced failure on self evaluation. Unpublished doctoral dissertation, Columbia University, 1954.

Underwood, B. J. *Psychological research.* New York: Appleton-Century-Crofts, 1957.

Zimmer, H. Self-acceptance and its relation to conflict. *J. consult. Psychol.*, 1954, **18**, 447–449.

Zuckerman, M., Baer, N., & Monashkin, I. Acceptance of self, parents, and people in patients and normals. *J. clin. Psychol.*, 1956, **12**, 327–332.

Zuckerman, M., & Monashkin, I. Self-acceptance and psychopathology. *J. consult. Psychol.*, 1957, **21**, 145–148.

(Received October 30, 1959)

Psychological Bulletin
1961, Vol. 58, No. 2, 122-131

THE CONSTRUCTION OF UNIDIMENSIONAL TESTS

JAMES LUMSDEN
University of Western Australia

It is the purpose of this article to review methods which have been suggested, either directly or indirectly, for the construction of unidimensional tests. No general survey of this topic appears to have been made previously but much help was obtained from critiques by Loevinger (1948), Guttman (1950a, 1950b,1950c), and White and Saltz (1957).

Definition of unidimensional tests. A unidimensional test may be defined simply as a test in which all items are measuring the same thing. A set of high jumps or a set of broad jumps is unidimensional. A mixture of high jumps and broad jumps is not. In psychological tests, however, items which appear to be of the same sort often turn out on closer investigation to be measuring different things so that this simple definition will not suffice for the construction of tests.

A more precise definition is given by considering the answer pattern that would be yielded by a unidimensional test with infallible items. If the items are arranged in order of difficulty placing the easiest first it will be found that a person who fails the first will fail all the other items; a person who passes the first and fails the second will fail all the subsequent items and so on. That is, the pattern of responses for five items could only be one of the forms shown in Table 1.

TABLE 1

PATTERNS OF RESPONSES FOR A UNIDIMENSIONAL TEST OF FIVE INFALLIBLE ITEMS

Total Scores	Item				
	1	2	3	4	5
0	F[a]	F	F	F	F
1	P	F	F	F	F
2	P	P	F	F	F
3	P	P	P	F	F
4	P	P	P	P	F
5	P	P	P	P	P

[a] *P* = Pass *F* = Fail.

With fallible items where the result may be affected by fluctuations in the ability of the subjects or in the difficulty of the items a perfect answer pattern may not be found even when the items do systematically measure the same thing. For our purposes, however, it is sufficient to take the answer pattern of Table 1 as providing a working definition of unidimensionality remembering that with fallible items the answer pattern will be disturbed by random error.

Criteria for evaluation. In evaluating the methods, major consideration will be given to the extent to which a method provides for:

1. A rational procedure for item selection
2. A criterion of unidimensionality
3. An index or measure of unidimensionality

A rational procedure for item selection is essential. Any method which provides no adequate indication of the most likely items to be discarded from the pool and which relies on a blind trial-and-error procedure to discover the unidimensional set of items will be hopelessly uneconomical for practical test construction. In general the method should be convergent so that the homogeneity of the item set increases as the procedure is applied and items are progressively removed from the original pool. Minor departures from this principle at critical stages (usually the beginning) are permissible so long as the number of trials to reach a convergent state of affairs is not too large.

A criterion of unidimensionality is necessary so that checks can be made from time to time and decisions made either to continue culling of the item pool or to stop culling because a homogeneous set of items has been obtained.

An index of the closeness of approximation to unidimensionality is also required. Failure to find a set of items which meets a strict criterion of unidimensionality is certainly possible and, indeed, very likely. The set of items in question may, however, be more unidimensional than any other measuring instruments available and would be preferable to a completely heterogeneous set of items alleged to measure the same attribute. The index of unidimensionality may be related to the procedure for selecting items and/or to the proposed criterion of unidimensionality, or, on the other hand, may be quite independent of either of these. It would be desirable for the sampling distribution of the index of unidimensionality to be known.

It should be noted that the index of unidimensionality is not quite the same as the measures of reproducibility discussed by White and Saltz (1957). Reproducibility as defined by White and Saltz confounds reliability and dimensionality since the measures are affected by random errors as well as by systematic differences in item content. An index of unidimensionality appropriate to the definition used here should be independent of random error.

Methods to be reviewed. Explicit consideration will be given only to classical item analysis, Loevinger's technic of homogeneous tests, the independence criterion method, Guttman's answer pattern method, and factor analysis. Most other methods are special cases of one or other of the listed methods and for the purpose of this review it is unnecessary to consider them. For example, criticisms of the Guttman procedure will apply also to the Cornell technique (Guttman, 1947a) and *H* technique (Stouffer, Borgatta, Hays, & Henry, 1952). Certain related techniques such as the Thurstone attitude scaling methods give tests of unidimensionality as a by-product but as test construction methods they are subject to the same criticisms as classical item analysis and the independence criterion.

Classical Item Analysis

Classical item analysis using an internal criterion attempts among other things to increase the average item-test correlation by selecting from the item pool those items which have the highest item-test correlation. It is well-known that this procedure tends to increase the homogeneity of the test.

From Table 1 it will be clear that for infallible items forming a unidimensional test the item-test correlation will be the maximum permitted by the shape of the distribution of test scores. With the answer pattern of Table 1 there is no overlap in the distribution of test scores for those who pass and those who fail a given item. The difference in mean test scores of passers and failers is thus a maximum and the biserial correlation between item and test is consequently maximized. It would appear then that if the culling of items proceeds to the point where the item-test correlations are all maximized the resulting test would be unidimensional. There are a number of difficulties which make this program unlikely to succeed.

With fallible items the maximum item-test biserial will not be reached. One solution would be to correct the obtained biserials for attenuation using estimates of the reliability of item and test scores. Accurate estimates of the reliability of a single item are not easily obtainable. Assuming that this difficulty can be overcome a test would be regarded as unidimensional if the biserial correlations between item and test approached the maximum after correction for attenuation.

Even granted the assumption that accurate estimates of item reliabilities can be obtained the method is not satisfactory. Consider the set of items with factor constitutions as follows:

$$x_1 = ma + nb + e_1$$

$$x_2 = ma + nb + e_2$$

$$x_3 = ma + nb + pc + e_3$$

$$x_4 = ma + nb + qc + e_4$$

$$x_5 = ma + nb + rc + e_5$$

where a, b, and c represent different orthogonal common factors, m, n, and p, q, r are loadings; and e_1, e_2, e_3, e_4, and e_5 are error factors.

Lumsden (1957) has shown that Items 1 and 2 form a unidimensional subtest and that Items 3, 4, and 5 with differing loadings on c are not unidimensional. Yet the method of maximizing item-test correlations will eliminate Items 1 and 2 first and no unidimensional test will be discovered. The only way out of this impasse would be to try sets at random which would make the procedure nonrational.

For this method the criterion of unidimensionality would be maximum biserial after correction for attenuation. No sampling distribution of corrected biserials appears to be available so that the significance of departures from the perfect fit cannot be assessed. This is specially important in this case since the estimates of item reliability on which correction is based are likely themselves to be quite unreliable.

The logical measure of unidimensionality would be average corrected biserial. This would need to be considered relative to the maximum obtainable biserial (biserial r has a maximum of 1.0 only when the continuous variable is normally distributed). A ratio of corrected biserial to its maximum similar to Loevinger's H_t suggests itself but the absence of a knowledge of its sampling distribution would restrict its value.

An obvious possibility would be to use the Kuder-Richardson Formula 20 with correction for variation in item difficulty suggested by Horst (1953). This statistic is, however, affected by random as well as systematic variance and is therefore, a measure of reproducibility rather than an index of unidimensionality. There would seem nothing to prevent the development of an index based on the ratio of obtained K-R 20 to the maximum K-R 20 for items with a given amount of random error.[1]

A search of the literature has not revealed any writer who has advocated the use of classical item analysis techniques as described above in order to produce unidimensional tests. Thorndike attempted to demonstrate the "homogeneity of intellect CAVD" by correlating scores on subgroups of

[1] I am indebted to John Ross (University of Sydney) for this suggestion.

items with scores on the total set of items and correcting the obtained r's for attenuation. Evidence was presented (Thorndike, Bergman, Cobb, Woodyard, 1926, p. 566) that these corrected correlations approximated 1.0 and Thorndike concluded that this demonstrated the homogeneity of CAVD tests. The logic of Thorndike's procedure is impeccable if applied to single items or to randomly selected subgroups of items but his subgroups were arranged so as to have, like the total set, equal numbers of Completion, Arithmetic, Vocabulary, and Directions items. Thorndike was thus merely able to show that the composite score obtained from his subsets was similar to the total score obtained from the complete set but not that the subsets or the total set were homogeneous in the sense used here. It is only fair to point out that Thorndike was mainly concerned to show that his easier sets of items and his harder sets gave the same sort of results as the total set.

Wherry and Gaylord (1943) suggest as an alternative to factor analysis an iterative procedure based on classical item analysis. In this procedure each item is correlated with total score; those items with the highest correlations are selected and a new total formed; all items (including those not selected in the first stage) are then correlated with the new total and the procedure is continued until a stable group of items is obtained. White and Saltz (1957) commend this method but it would not appear to avoid any of the difficulties of classical item analysis.

Loevinger's Technic of Homogeneous Tests

Loevinger's procedure is closely related to classical item analysis and indeed she indicates (Loevinger 1947, p. 26) that the earlier work by Thorndike on the CAVD tests may have been influential in the development of her procedure.

The procedure is based on two statistics: the "homogeneity of an item with a test" and the "homogeneity of a test." The first of these is to be used as a tool for item selection and is a development of Long's (1934) index of overlapping. The formula for this is given by Loevinger as:

$$H_{it} = 1 - \frac{2\,(\text{“passes” below or tied with “fails”})}{PQ - \text{“passes” one above “fails”}}$$

where P is the number passing the item and Q is the number failing the item. It is clear that for a perfectly unidimensional test as defined by Table 1 H_{it} will equal 1.0 since there will be no subjects who pass an item who will have scores below or tied with subjects who fail the item. Using this statistic to cull a mixed set of items will, however, be subject to all the difficulties encountered with classical item analysis.

The index of unidimensionality is provided by the "homogeneity of a test," H_t. Loevinger notes that for a perfectly heterogeneous test $p_{i/j} = p_i$ (i.e., probability of passing an Item i having passed another Item j is the same as the overall probability of passing Item i). For a perfectly homogeneous test as defined by Table 1, $p_{i/j} = 1.0$ for all cases where $p_i > p_j$ (i.e., where Item i is easier than Item j). From this it will be seen that $p_{i/j}$ has a minimum value of p_i for all cases where $p_i > p_j$.

Loevinger then considers the sum:

$$S=\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} p_j(p_{i/j}-p_i)$$

where m is the number of items and the item pairs are all such that $p_i > p_j$.

This sum will have a maximum value given by

$$S_{\max}=\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} p_j(1-p_i)$$

for a perfectly homogeneous test and a value of zero for a perfectly heterogeneous test. To provide an index with the formal properties of a minimum of zero and a maximum of 1.0, Loevinger divides S by $S_{\max}$ to give:

$$\frac{S}{S_{\max}}=\frac{\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} p_j(p_{i/j}-p_i)}{\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} p_j(1-p_i)}$$

Loevinger provides a formula for estimating H_t from sample statistics but points out that the sampling distribution is unknown and that the estimate is not even known to be unbiased.

Interdependence Criterion

Lazarsfeld (1950), Tucker (1952), and Lord (1952) have pointed out that with a unidimensional test the probability of success on one item is independent of success in any other item for subjects with the same true score. This is at first sight paradoxical because it would seem obvious that items which are measuring the same thing should be highly correlated. But when only subjects of the same true ability are considered then items which are measuring this ability and nothing else can differ only through error and will exhibit no systematic variance. If we take subjects who are exactly 6 feet tall then different measures of height will vary only through error so that the measurements will be independent, uncorrelated. The independence criterion is undoubtedly valid and is more general than any other. It makes no assumptions about the distribution of ability or rectilinearity of regression.

The criterion suggests a procedure for constructing unidimensional tests. It would be possible to obtain results on a pool of items from a large group of subjects, to choose a number of subjects with the same total score, and then to determine say by χ^2 whether the items are independent or not. If certain items turned out not to be independent these could be rejected, new totals worked for all subjects in the original group, a new group with the same total score determined, and the χ^2 test repeated.

The true scores on the test are not, however, known and the estimates obtained from the raw test scores are not satisfactory. O'Neil (1954) has shown that, for subjects with the same obtained score, items, even if unidimensional, tend not to be independent but to be negatively correlated. If there are only two items for example then for subjects with an obtained score of 1 the items have a tetrachoric correlation of -1.0 since if a subject has the first item right he must have the second one wrong and vice versa. In mathematical terms $p_{i/j}=0$ instead of p_i as required by the independence criterion. This effect is known to decrease as the number of items is increased and it is possible that the independence criterion may be workable for fairly large groups of items. With infallible items there is, of course, no problem since true scores will then equal the obtained scores and the quoted example could not occur (if the items were unidimensional).

Even if this problem is overcome, the culling of items is likely to prove arduous. All items in the pool are likely to be correlated on the first trial. In the absence of any knowledge about the number of items in the unidimensional set it is impossible to say whether the unidimensional items will be more or less intercorrelated than the items it is desired to reject. No rational, convergent procedure of item culling is available using the independence criterion.

No special index of unidimensionality is suggested for this method. This, of course does not matter, since if the method was otherwise suitable an index could be borrowed from one of the other methods.

Answer Pattern Methods

The Guttman procedure (1944) is the most important of the answer pattern methods and is the only one discussed here. Some earlier writings by Walker (1931, 1936, 1940) and Ferguson (1941) have the first explicit discussions of the relationship between answer pattern and other test characteristics but no suggestions for test construction were made.

The answer pattern procedure consists essentially of inspecting the answer pattern and removing items so that the remaining items have patterns which are as near as possible to those of Table 1. It is clear that for infallible items this procedure could be easily carried out and that a simple inspection of the answer patterns would provide a clearcut criterion of unidimensionality. For items which exhibit slight departures from unidimensionality the procedure would be to eliminate items until the closest possible approximation consistent with retaining a sufficient number of items was obtained. For this, some measure of the closeness of approximation to unidimensionality is required. Guttman uses the coefficient of reproducibility which is the proportion of responses which can be correctly predicted from the total raw score. For a perfectly unidimensional test it will be seen from Table 1 that the reproducibility coefficient will have the value 1.0. Guttman suggests that a test may be regarded as a "scale" (i.e., as unidimensional) if the coefficient of reproducibility exceeds .90.

The coefficient of reproducibility has been criticised severely by Festinger (1947) and Jackson (1949) because it does not allow for the chances of obtaining high values when the items are heterogeneous (e.g., with only a few items of widely differing difficulties). Guttman (1947b) replied to criticism claiming that such factors as the number of answer categories and the range of difficulty were taken into account before calculating the coefficient of reproducibility. Guttman does not give explicit rules but improvements to the reproducibility coefficient have been suggested by Jackson (1949) and Green (1954) which overcome some of the problems.

The reproducibility coefficient, however modified, does not permit of a distinction being made between random and systematic scale discrepancies. Guttman claims (1950a, 1950b, 1950c) that the distinction may be made by examining the patterns of scale discrepancies and presents tables (p. 161) which purport to represent scale patterns for a perfect scale, a scale with random error, and a scale with systematic error. Evidence for random error in an item is said to be provided when scale errors are distributed randomly around the cutting point for the item; evidence for systematic error when the scale errors are grouped in a systematic fashion. While this claim is undoubtedly correct (such systematic groupings are the basis of all the statistical analyses proposed for the problem) it is difficult to see how these groupings may be discovered by inspection and distinguished from random errors when the random errors are fairly large.

Guttman (1950b) has explicitly denied any intention to use scale analysis for the selection of items. His scalogram was designed merely to discover approximate cutting points for attitude scale items. Guttman indeed claims that the task of scale analysis is to discover scales rather than to construct them and states that if a universe of attributes is scalable then any subset of items from that universe is scalable. Item culling is by this argument unnecessary. The difficulty is that a test constructor (or discoverer) does not know precisely what "universe of attributes" he is sampling. Without precise definition he may sample a number of related universes. Item culling procedures are designed to distinguish between groups of items selected from different universes.

It may be seen then that the answer pattern method provides no rational culling plan for use with fallible items. The index of unidimensionality provided by the plan is the coefficient of reproducibility which, despite improvements on the early Guttman form, does not distinguish between systematic and random error.

Factor Analysis

It is difficult to give due credit to whoever first suggested the use of factor analysis in the construction of unidimensional tests. The idea is sufficiently obvious to be thought at least implicit in the writings of Spearman, Thurstone (1947), and other early factorists. The factor analyses of test items by McNemar (1942), Burt and John (1943), and others clearly suggest it. Papers on related topics by Ferguson (1941), Wherry and Gaylord (1943, 1944), Carroll (1945), and Loevinger (1948) discuss with varying degrees of completeness the possibility of factor analyzing items in test construction.

Under restrictions which appear plausible for ability test items it is easy to show (vide Lumsden, 1957) that for a unidimensional test the matrix of tetrachoric item intercorrelations is of unit rank. One factor analytic procedure for constructing unidimensional tests is to extract a single factor from the item intercorrelations, cull out the items which have large residuals, reanalyze, and continue until a satisfactory fit to a single factor solution is obtained. Wolfle (1940) in a well-known jibe at Brown and Stephenson (1933) said: "if one removes all tetrad differences which do not satisfy the criterion, the remaining ones do satisfy it" (p. 9). That is exactly what is done in this factor analytic technique of constructing unidimensional tests. The difference between the two situations is, of course, that Brown and Stephenson had asserted that their tests, all of them, would meet the tetrad difference criterion, while here it is merely hoped that a subset of items will meet the criterion.

The procedure is quite simple. But is the culling procedure rational? Will the set of items converge to unidimensionality?

It is evident that convergence of the factor analytic procedure to a unidimensional subset of items cannot be guaranteed. If the unidimensional set is much less numerous than the heterogeneous items in the pool then it is probable that the unidimensional set will not have sufficient influence on the nature of the first factor extracted to prevent the occurrence of large residuals among the unidimensional set. These items will be discarded first and the procedure will not converge to a single factor solution.

If, however, the items are carefully preselected on empirical and a priori grounds, it seems likely that the state of affairs of the preceding paragraph will not occur. If items are deliberately made parallel or if there is evidence for parallelism then it would follow that the dimension of any unidimensional test and the dimensions of the heterogeneous items in the total pool, will normally be highly correlated. In this circumstance the influence of the unidimensional set on the first factor extracted may well be greater than the actual numbers of items suggest, and the method may therefore be expected to converge. The procedure of preselecting will also tend to increase the size of the unidimensional set in the pool and this will also increase the probability of convergence.

Lumsden (1959) found that four subsets of number series items selected on a priori grounds converged rapidly and that three of them met a fairly stringent test of unidimensionality when cross-validated with a fresh group of subjects.

One procedure that should almost guarantee convergence (if a sizable unidimensional set exists) is to carry out a preliminary complete centroid analysis and then to select for further analysis those items which appear in narrow strips (i.e., roughly co-linear) in the factor space. This appears to be the procedure advocated by Cattell (1957) for his factor homogeneous scale except that he would require the additional restriction that the factor have significance in a more general factor space than that provided by the item intercorrelations. The complete centroid procedure with rotation could indeed be used without further analysis except that the problems of estimating communalities and determining goodness of fit are more complicated than for the unit rank case.

The criterion of unidimensionality suggested for item culling is the size of the residuals. This must be considered with relation to the sampling distribution of residuals. Unfortunately there is no exact solution to this problem. Many methods have been suggested (Cattell, 1952) but none can be regarded as satisfactory. A reasonable solution for test construction purposes would be to use one of the simpler procedures (e.g., standard error of average r) and apply it rather severely. Increased availability of automatic computing services may permit the use of maximum likelihood methods of factorizing which provide a test for rank.

An index of unidimensionality appropriate to the method is the ratio of first factor variance to total bipolar factor variance after a complete centroid analysis with subjects who were not used for item selection. In most cases the ratio of first to second factor variance would seem to give a reasonably useful index. This index has no fixed maximum value and little is known about the extent to which it may be affected by errors of sampling or of measurement.

DISCUSSION

It seems clear that none of the methods examined can be regarded as satisfying all three of the main criteria. Only factor analysis appears to offer a rational procedure for item selection. The criteria and indices of unidimensionality are unsatisfactory for all methods.

This review has considered each of the methods as if they were complete, self-consistent creations of a single writer. With the exception of the Guttman answer pattern method and the Loevinger method this is not so. The various "natural" criteria and indices suggested for each of the methods are not necessary consequences of the choice of item selection method. Combinations of different elements from different methods are possible and this circumstance justifies a modified optimism. Thus a modification of the coefficient of reproducibility which produced an acceptable index of unidimensionality would not be cogent evidence for adopting an answer pattern method but would greatly improve all methods.

Greatest emphasis has been deliberately placed on item selection rationale since this topic appears to have been relatively neglected in the literature of the problem. Great advances appear unlikely unless the development of criteria and indices of unidimensionality is closely related to item selection procedures.

Summary

Five methods of constructing unidimensional tests (classical item analysis, Loevinger's procedure, the independence criterion method, the answer pattern method, and factor analysis) have been considered with respect to their provision for: a rational procedure for item selection, a criterion of unidimensionality, and an index of unidimensionality.

It has been argued that only factor analysis provides a rational procedure for item selection. No method has a fully satisfactory criterion of unidimensionality. The index of unidimensionality suggested for the factor analytic method is the ratio of first to second factor variance. This suffers from the absence of any knowledge of sampling fluctuations, but this weakness is shared by the only reasonable alternative, the coefficient of reproducibility.

REFERENCES

Brown, W., & Stephenson, W. A test of the theory of two factors. *Brit. J. Psychol.*, 1933, **23**, 352–370.

Burt, C., & John, E. A factorial analysis of the Terman-Binet tests. *Brit. J. educ. Psychol.*, 1943, 12, 156–161.

Carroll, J. B. The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika*, 1945, **10**, 1–19.

Cattell, R. B. *Factor analysis*. New York: Harper, 1952.

Cattell, R. B. *Personality and motivation structure and measurement*. New York: World Book, 1957.

Ferguson, G. A. *The reliability of mental tests*. Univer. London Press, 1941.

Festinger, L. The treatment of qualitative data by "scale analysis." *Psychol. Bull.*, 1947, **44**, 146–161.

Green, B. F. Attitude measurement. In G. Lindzey (Ed.), *Handbook of social psychology*. Cambridge, Mass.: Addison-Wesley, 1954.

Guttman, L. A basis for scaling qualitative data. *Amer. sociol. Rev.*, 1944, **80**, 139–150.

Guttman, L. The Cornell technique for scale and intensity analysis. *Educ. psychol. Measmt.* 1947, **7**, 247–279. (a)

Guttman, L. On Festinger's evaluations of scale analysis. *Psychol. Bull.*, 1947, **44**, 451–465. (b)

Guttman, L. The basis for scalogram analysis. In S. A. Stouffer (Ed.), *Measurement and prediction*. Princeton: Princeton Univer. Press, 1950. (a)

GUTTMAN, L. The problem of attitude and opinion measurement. In S. A. Stouffer (Ed.), *Measurement and prediction.* Princeton: Princeton Univer. Press, 1950. (b)

GUTTMAN, L. Relation of scalogram analysis to other techniques. In S. A. Stouffer (Ed.), *Measurement and prediction.* Princeton: Princeton Univer. Press, 1950. (c)

HORST, P. Correcting the Kuder-Richardson reliability for dispersion of item difficulties *Psychol. Bull.*, 1953, **50**, 371–374.

JACKSON, J. M. A simple and more rigorous technique for scale analysis. In, *A manual of scale analysis.* McGill University, 1949. (Mimeo)

LAZARSFELD, P. F. The logic and mathematical foundation of latent structure analysis. In S. A. Stouffer (Ed.), *Measurement and prediction.* Princeton: Princeton Univer. Press, 1950.

LOEVINGER, JANE. A systematic approach to the construction and evaluation of tests of ability. *Psychol. Monogr.* 1947, **61**(4, Whole No. 285).

LOEVINGER, JANE. The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychol. Bull.*, 1948, **45**, 507–529.

LONG, J. A. Improved overlapping methods for determining the validities of test items. *J. exp. Educ.*, 1934, **2**, 262–268.

LORD, F. M. A theory of test scores. *Psychometric Monogr.*, 1952, No. 7.

LUMSDEN, J. A factorial approach to unidimensionality. *Aust. J. Psychol.*, 1957, **9**, 105–111.

LUMSDEN, J. The construction of unidimensional tests. Unpublished master's thesis, University of Western Australia, 1959.

MCNEMAR, Q. *The revision of the Stanford-Binet scale.* Boston: Houghton Mifflin, 1942.

O'NEIL, W. M. A problem of testing a set of items for unidimensionality. Paper read at British Psychological Society (Australian Branch), Perth, 1954.

STOUFFER, S. A., BORGATTA, E. F., HAYS, D. G., & HENRY, A. F. A technique for improving cumulative scores. *Public opin. Quart.*, 1952, **16**, 273–291.

THORNDIKE, E. L., BERGMAN, E. D., COBB, M. V., & WOODYARD, E. *The measurement of intelligence.* New York: Teachers' College, Columbia Univer., 1926

THURSTONE, L. L. *Multiple factor analysis.* Chicago: Univer. Chicago Press, 1947.

TUCKER, L. R. A level of proficiency scale for a unidimensional skill. *Amer. Psychologist*, 1952, **7**, 408. (Abstract)

WALKER, D. A. Answer-pattern and score scatter in tests and examinations. *Brit. J. Psychol.*, 1931, **22**, 73–86.

WALKER, D. A. Answer-pattern and score scatter in tests and examinations. *Brit. J. Psychol.*, 1936, **26**, 301–308.

WALKER, D. A. Answer-pattern and score scatter in tests and examinations. *Brit. J. Psychol.*, 1940, **30**, 248–260.

WHERRY, R. J., & GAYLORD, R. H. The concept of test and item reliability in relation to factor pattern. *Psychometrika*, 1943, **8**, 247–269.

WHERRY, R. J., & GAYLORD, R. H. Factor pattern of test items as a function of the correlation coefficient: Content, difficulty and constant error factors. *Psychometrika*, 1944, **9**, 237–244.

WHITE, B. W., & SALTZ, E. Measurement of reproducibility. *Psychol. Bull.*, 1957, **54**, 81–99.

WOLFLE, D. Factor analysis to 1940. *Psychometric Monogr.*, 1940, No. 3.

(Received January 4, 1960)

Psychological Bulletin
1961, Vol. 58, No. 2, 132-142

CONTRIBUTIONS AND CURRENT STATUS OF *Q* METHODOLOGY

J. R. WITTENBORN
Rutgers University

For the present purposes we shall take our definition of *Q* methodology from Stephenson (1953), not because he has been succinct, but because he has been more comprehensive in his interest than other writers. Accordingly, *Q* as conceived by Cattell or Mowrer and related methodological topics discussed by such writers as Cronbach are not included in the present review of recent studies.

In his 1953 publication, *The Study of Behavior*, Stephenson informs us with a modesty which is characteristic for this book that " . . . the science of behavior can be immeasurably improved by attending to a few principles upon which we have based the method now well known as '*Q*-technique' " (p. 1). Time does not permit the long series of quotations which would be necessary in order to indicate fully Stephenson's concept of *Q* methodology, but a few additional quotations remind us that he was definite in his point of view.

> Our object has been to make it possible for studies to be undertaken on single cases (p. 2).[1] Briefly, a statement of the kind "All crows are black" is a general proposition. To say that "A crow is black" is clearly singular, but not testable. When, however, we can point to a particular crow X and assert that *it* is black, a singular testable proposition is at issue (p. 42). There never was a single matrix of scores to which *both* R and Q apply (p. 15).

A defining summary of Stephenson's view of *Q* methodology could include at least six points:

1. *Q* method appears to require ipsative variables, particularly *Q* sorts.

2. *Q* method lends itself to correlations between people or between different conditions for the same person.

3. *Q* method requires a conceptually structured set of statements in order to interpret the correlations between people—each set of statements comprising systematic combinations of different levels of the various hypothetical effects.

4. *Q* method permits a study of a person by means of analysis of variance of the statements, assuming that the sorted statements were initially structured as replications of the possible combinations of a priori effects and levels of reaction.

5. *Q* method favors a dependency type emphasis in factor analyses with rotations determined by the nature of the propositions concerning the variates.

6. *Q* method leaves unanswered the question of the parent population from which the individual is drawn; the method examines singular propositions on the assumption that somewhere there are more people like the one under scrutiny.

To date the most conspicuous use of *Q* methodology has been made by the so-called "self psychologists" who view discrepancies between one's self-perception and the perception of an ideal self as an indication of

[1] "By a 'single case' we mean, for the moment, a single person under study or a single group of interacting persons. . . . what is involved is whether individual differences are postulated or whether singular propositions are being tested. The latter alone are our concern" (Stephenson, 1953, p. 2).

maladjustment. This interpretation of psychological maladjustment is consistent with Rogers' belief in the self-actualizing function of the personality and, as a consequence, finds direct application in his studies of the efficacy of psychotherapy (Rogers & Dymond, 1954). The work done by the Chicago group during the first half of the present decade used the now familiar device of correlating the individual's *Q* sort describing his true self with his *Q* sort for his ideal self. Increases in these correlations during the course of therapy were taken as evidence of improvement. This was a notable application of *Q* methodology because in the opinion of many persons these studies comprised the first acceptable indication that psychotherapy was efficacious in producing personality change. Beyond using *Q* methodology in establishing this important landmark, the Chicago group was able to illuminate some of the features of the psychotherapeutic process by factor analyzing the intercorrelations among various *Q* sorts for a given patient. The case of Mrs. Oaks is illustrative. During the course of therapy her concept of self became much more favorable, and there were some changes in her concept of an ideal self. Her therapeutic progress was summarized by a factor analysis of the intercorrelations among *Q* sorts made at various stages in the course of her therapy.

There have been several evaluations of the validity of *Q* sorts as evidence of adjustment, particularly from the standpoint of their appropriateness as criteria for therapeutic change. For example, Friedman (1955) supported the self-ideal consistency concept of good adjustment by a study which involved a comparison between normals and neurotics. The neurotic group was described as tending to regard their self-qualities as very much different from the way they would like them to be. Cartwright (1957) emphasized the consistency interpretation of good adjustment by showing that after psychotherapy subjects describe themselves in relation to important persons in their environment with as much consistency as controls. An increase in self-ideal congruence for a group of high school boys after counseling was reported by Caplan (1957), and Turner and Vanderlippe (1958) reported that college students with high self-ideal congruence tended to have more extracurricular activities and to have higher scholastic averages than students with low self-ideal congruence.

The apparent validity of self-ideal congruence was examined by Chase (1957), who compared adjusted and maladjusted hospital cases with respect to the various possible correlations involving *Q* sorts. Only those correlations containing the self-sort distinguished between the adjusted and the maladjusted group.

The *Q* sort approach to adjustment is subjected to further scrutiny by Kogan, Quinn, Ax, and Ripley (1957). Using two comparable samples, one psychiatric patients and the other university students, sorts for four different conditions were obtained. The average sorts for the patient and the student groups were correlated for each of the four conditions. It was found that a great portion of the variance in these correlations could be accounted for in terms of either of two extraneous variables, the social desirability of the sorted statements or a sickness-health variable. Edwards (1955) had described the importance of social desirability in *Q* sorts as early as 1955.

There are reports which challenge the appropriateness of *Q* sorts as a direct evidence of the efficacy of psychotherapy. For example, Taylor (1955) undertook a study on the assumption that repeated introspection would produce therapeutic type changes in self-concept. His subjects made repeated *Q* sorts. In consequence of this procedure, there was an increase in positive self-concepts, and the self-ideal correlation increased. From this one need not infer that self-concepts are unstable, however. Engel (1959) examined the self-concepts of a group of adolescents in 1954 and again in 1956, and reported that items relative to positive self-concepts had appreciable stability as indicated by a stability correlation of .69.

Levy (1956) challenged the meaning of self-ideal discrepancies by comparing self-ideal correlations based on the Butler and Haigh (1954) items with the correlation between sorts for an actual and an ideal home town. He found these two sets of actual-ideal correlations to be correlated with each other to the order of .70. Because of this, he suspects that the discrepancies perceived between actual and ideal states of affairs have implications for the individual's adjustment, regardless of the area in which the discrepancy is shown.

Although *Q* sorts are frequently employed in the published literature, the use is often relatively uncritical. In some instances it would appear that a normative procedure would have served the investigator's purposes better than the ipsative *Q* sorts. It appears probable, however, that investigators have been encouraged by the availability of *Q* sort procedures, and some of the resulting studies might not have been undertaken if only a normative type emphasis had been available. For example, in Stewart's (1958) study of the relationship between manifest anxiety and mother-son identification, facets of mother-son identification are readily revealed and usefully quantified by correlations between various sorts provided by the mothers and their sons. Stewart found that the boys with the greatest manifest anxiety were those with the greatest discrepancy between their self-perceptions and their mother's ideal for them.

Correlating *Q* sorts was a convenient device for Kalis and Bennett (1957), who wished to show that communication between the patient and members of his family was improved for those patients whose hospitalization had been successful. The importance of similarity of self-perceptions in interpersonal relationships is further illuminated by Corsini's (1956) use of *Q* sorts in his study of happiness in marriage. These studies are reminiscent of a report by Revie (1956), who used *Q* sorts to describe both the teacher's and the school psychologist's concept of pupils. As a result of their interaction, both the teacher's and the psychologist's concept of the pupil changed.

Q sorts have been used in many different ways, particularly in the study of personality. Shontz (1956) used *Q* sorts in order to examine the concept of a healthy personality, while Reznikoff and Toomey (1958) worked out a system of weightings whereby observers' *Q* sorts of patients may be scored to estimate the degree of emotional disturbance. Epstein and Smith (1956) used *Q* sorts as a sociometric device by having students *Q* sort their fellows with respect to the degree of hostility in their behavior. Fiske and Van Buskirk (1959) used *Q*-sort procedures in order to examine the stability of sentence completion test interpretations, and Doleys and Kregarman (1959) report that self-ideal congruence does not measure frustration tolerance. Nahinsky (1958) used a self-ideal comparison to distinguish career from noncareer naval officers, and Whiting (1959) had nurses, patients, and physicians sort statements concerning the importance of various aspects of the nurse's work. This appears to be the kind of a study where rating scales, inventories, or check lists could not have served the investigator's purposes as well as the ipsative sort.

The unique value of *Q* sorts has not been made sufficiently explicit to permit an investigator to know the kinds of situations which call for ipsative procedures and the kinds of situations in which his purposes will be better served by normative procedures. There are numerous studies in the literature which employ *Q* sorts without indicating why this particular method was chosen. Sometimes it appears that *Q* sorts are used because no reliable normative instrument is available to distinguish between persons along a relevant continuum. The question of the reliability or the validity of the *Q* sort is rarely raised, and if practice alone were considered, one could infer that reliable and valid ipsative distinctions based on a *Q*-sort procedure are much easier to establish than reliable and valid normative procedures. Even if this were true, and your reviewer has not seen material which would lead to this conclusion, one could still question the use of an ipsative procedure showing intraindividual differences when a normative procedure dealing with interindividual differences appears to be indicated by the general requirements of the investigation. For example, Morsh (1955) described the use of a *Q*-sort procedure in securing the classes' evaluation of the teachers. Why an ipsative type evaluation of teachers is preferable to a normative-type procedure is not indicated in his report.

In a study of the relationship between some personality variables and speed in decision making, Block and Peterson (1955) used the staff's *Q* sort of the subject as a measure of personality. Although the results of this investigation are interesting and worthwhile, it would appear that the emphasis is one of individual differences and that a normative measure of personality would have been the logical choice. Both Cattell (1944) and Guilford (1954) have warned that ipsative measures should not be used in attempts to study individual differences. The amount of error in such a maneuver need not be invariably great, however. For example, Block (1957) matched items from a *Q* sort with items that were used in a normative-type rating and found that in one sample the correlations between various items ranged from .63 to .88, while similar correlations for another sample ranged from .31 to .74. Apparently, the error involved in using ipsative item scores in a normative manner may vary greatly from item to item and from sample to sample.

In addition to their applications in various studies of personality, *Q* sorts are also applied in the study of psychopathology. For example, Rogers (1958) found that the self-ideal congruence for paranoid schizophrenics was greater than for normals. His approach is noteworthy because of its novelty. Instead of having his subjects sort cards, he asked them to manipulate a red square over a blue square, with the red square representing the self, the blue square representing the ideal, and the overlap representing the degree of congruence. Although the spatial interpretation that the subject gives his judgment is absolute and could lend itself to normative treatment, the sense of these manipulations is clearly ipsative. This study, which was published in 1958, showed a high degree of self-ideal congruence for paranoids and can be compared with Friedman's 1955 study which included a sample of paranoid schizophrenics. Friedman found that only 3 of his 16 paranoids showed a low self-ideal correlation.

Other schizophrenics are much more distinctive with respect to their behavior in the *Q*-sort situation. For example, Helfand (1956) asked various subjects, including schizophrenics, to simulate the *Q* sort of a former patient whose autobiography they read. He then computed the correlations between the sorts provided by his various subjects and the sort provided by the former patient. He found that the schizophrenics' simulated sorts had the lowest correlations of all. He ascribes this to a limitation in role-taking ability. A recent paper by Fagan and Guthrie (1959) tells us more about the schizophrenics. Their subjects were asked to describe themselves in one sort and to describe an average person in another. The subjects were intercorrelated for the two different sorts, and the two sets of intercorrelations were factor analyzed. The authors concluded that schizophrenics, like many other patients, view themselves differently from the way they view other persons.

The mothers of schizophrenic patients have also been studied by *Q* methodologists. Shepherd and Guthrie (1959) had the mothers of 20 male schizophrenics sort 100 items concerning children and family life. These sorts made it possible to correlate each mother with every other one. The correlations were factor analyzed and the resulting factors—identified as Detached Authoritarianism, Inadequacy and Inconsistency, Pervasive Control, Sophisticated Denial of Inadequate Mothering, and Annoyance and Rejection—broaden our view of the various qualities or dimensions of schizophrenic mothering. One immediately becomes interested in the manner in which one might generalize from these mothers of schizophrenics to other mothers and thereby gauge how broadly applicable one might find such dimensions of schizophenic mothering. Unfortunately, this is one of the ways in which *Q* methodology is weakest. We do not know what population the individual or individuals under scrutiny represent. Stephenson (1953) seems to feel that this really does not matter as long as he can assume that there are other similar individuals somewhere. He calls ducking this practical issue testing a "singular proposition."

There are many published studies which involve factor analyzing intercorrelations between persons. According to Stephenson's criteria, however, only a few of these would qualify as an application of *Q* methodology. Since Stephenson states that there is no matrix of correlations which can be studied by both *R* and *Q* methods, one is inclined to conclude that one can appropriately intercorrelate persons for *Q* purposes only when the similarity of the persons is expressed by a correlation based on ipsative scales, i.e., scales on which people have distinguished between items and not necessarily scales which distinguish between people on any normative basis of individual differences. Thus, your reviewer's factor analyses of various diagnostic groups, although designed to show that different varieties of patients may have the same diagnosis, should not be considered as application of *Q* methodology because the correlations between persons were based on standard rating scales designed to show individual differences. There are many such obverse factor analyses, and although they are commonly called *Q* studies, they do not meet Stephenson's criteria. The Bendig and Hamlin (1955) investigation of Rorschach scoring categories is another study of this type.

Perhaps the most valuable applications of factor analysis in *Q* methodology may come from studies of therapeutic phenomena. The possibilities of such an approach were anticipated as early as 1951, when Fiedler published a factor analytic study of differences between therapists from different schools and with different levels of training. Despite the potential of such studies for helping to place psychotherapy on a rational, empirically verifiable basis, only a few students of psychotherapy appear to be ready to study therapeutic phenomena with the systematic planfulness which *Q* methodology could facilitate. In one such study, Nunnally (1955b) had a therapist describe a patient by means of *Q* sorts on eight successive occasions. The factor analysis of these intercorrelations yielded two factors—one concerning relationships with the therapist and the other relating to intrapersonal confidence.

The Peterson, Snyder, Guthrie, and Ray (1958) investigation of therapeutic biases provides a promising exploration. They approached their study in a sound manner by thoughtfully structuring the sample of statements which comprised their *Q* sorts, systematically using variations of such hypothetical dimensions as direction of gain, attitudes, mode of change, and area of conflict. The sample of therapists who were intercorrelated was drawn from graduates of their own program so that one is not left up in the air with respect to the population of persons to whom the results may be generalized. As is usual in such studies, the factors were interpreted on the basis of the items which received a characteristic sort by persons who had high loadings on the factor. The practice of interpreting persons in terms of item smacks of *R* methodology and reminds us that people are usually more distinguished by their behavior than behavior is distinguished by the people who perform it.

Thrush published an interesting study in 1957. Using a sample of 60 statements descriptive of problems encountered by a counseling agency, the staff made sorts of the level and kind of service each problem would require. These sorts were made in 1952 and again in 1956. On the basis of these sorts, the members of the staff were intercorrelated for each of the years and the two sets of intercorrelations were factor analyzed separately. A comparison of the results indicated that the emphasis in the agency had shifted from vocational counseling to personal adjustment counseling. Although studies of this kind are illuminating, they remind us that we have no rigorous basis for comparing the results of factor analyses to test an exact statistical hypothesis. The question of how one should generalize from a

Q-type study is usually disregarded. Conger, Sawrey, and Krause (1956) point to an aspect of this problem in their study of Beck's "The Six Schizophrenias" (1954).

In commenting upon factor analysis in *Q* methodology, one should remember that Stephenson indicated that the correlations should be in part expressive of the *effect* of different kinds of operations. He intended that the intercorrelated variates should, in some manner or another, be regarded as dependent variables in an experimental sense and not merely descriptive dimentions of a static situation. From the standpoint of this emphasis, the Sweetland and Frank (1955) study of ideal psychological adjustment is not a good example of *Q* methodology because its purpose appears to be to describe kinds of psychological adjustment rather than to reveal the effects of certain operations, i.e., it is not a dependency-type analysis. This descriptive use of the *Q*-type factor analysis is not unique to Sweetland and Frank, however; other examples would include Broen's (1957) factor analytic study of religious attitudes.

Many of the samples of statements which have been sorted in *Q* methodology appear to have been somewhat informally assembled, and as a consequence, the analyses performed on the sorts provided by various persons or by the same person under various instructions have an uncertain meaning. We do not know from what parent population of behavior they might conceivably be drawn or from what specific theory they could have been generated. It is probably for this reason that we find relatively few studies where the *Q* sort arrays for an individual or a group of individuals are submitted to an analysis of variance. This is unfortunate because the difference or similarity between the *Q* sorts of two individuals or the *Q* sorts of an individual under two or more conditions must be explained in terms of the sorted items which comprise the *Q* arrays. If the items had been included in the sample as a priori representatives of theoretically relevant classes of behavior, then the order given to the items could in the case of any given *Q* sort be entered into an analysis of variance. In this way the relative status which the sorter assigned to various a priori classes of items could be revealed. In many studies, however, a defensible a priori classification of behavior with respect to kinds and levels is not possible because the area of inquiry is not well known, no systematic theory can be confidently applied, and in a sense the investigation is exploratory. If, in the study of such an area of behavior, *Q* methodology were indicated, it would seem desirable first to intercorrelate and factor analyze the items in the *R* tradition. Then a sample of statements for *Q* sorts could be arranged so that the various factors could be represented in a balanced design. From such structured samples of statements, the *Q* methodology could be applied in the recommended manner by first factor analyzing the variates (e.g., people) and then explaining the *Q* factors in terms of an analysis of variance of the sorts provided by the variates. The reviewer saw no studies where the domain of behavior was first explored by an *R*-type analysis as a basis for building a structured sample of statements for the *Q* sorts. Where analysis of variance had been applied to *Q* sort arrays, the investigator had carefully structured his sample on an a priori basis. Such studies are few and tend to be found in the recent literature.

One of the earlier studies involving an analysis of variance was provided in 1956 by Kerlinger who constructed a set of *Q* statements which represent two kinds of educational attitudes interpreted at four different levels each. The levels for each class were then systematically replicated with 10 statements each, so that there were 80 statements in all.

In 1958, Rawn published a study of transference and resistance in psychotherapy. The statements to be sorted conformed with the requirements of a balanced block design involving two levels of resistance and three classes of transference. These categories of class and level could be combined to form six kinds of statements. Each of these types of statements was interpreted in 15 different ways to form the replications, and accordingly the structured sample comprised 90 statements in all. These statements were sorted by different raters and for different sessions of recorded psychotherapy. Because of the way the sample was structured the investigator could perform an analysis of variance for the various sorts as well as factor analyze the intercorrelations among the sorts. His purposes required the analysis of variance only, however.

Perhaps some of the most substantial values to accrue from the point of view known as *Q* methodology may lie in the fact that more of us have become increasingly thoughtful about many matters which we had formerly disregarded or postponed. Possibly one of these neglected matters is the hiatus between the clinicians who continue to be interested in intra-individual differences and the psychometricians who, acting from the standpoint of interindividual concepts of reliability, have dismissed intra-individual differences as trivial or of no possible consequence.

There is a general tendency for investigators to compute correlation coefficients without giving much thought to the meaning or the determiners of the relationship. *Q* methodology is leading us to think more realistically about features which contribute to the degree of correlation between either subjects or items. If, for example, the sample of items is not homogeneous, it would seem possible for several pairs of persons to be equally correlated with each other but for the various pairs to have their respective correlations because of different items of behavior. As a consequence, none of the items may characterize all of the intercorrelated persons. Presumably a similar kind of situation could exist if items were intercorrelated for a group of persons representing subsamples of different populations. In such a case the correlations found between any two items might vary considerably if they were separately calculated for the various subsamples instead of being calculated for the heterogeneous group. Obviously, the investigator is on shaky ground when he assumes that a correlation based on one sample is descriptive of some other sample which is comprised in some different manner. The composition of the sample with respect to persons can obviously influence the correlation between items, or the composition of a sample with respect to items could influence the correlation between persons.

Some aspects of this problem of subject homogeneity were discussed by Block in 1955, and in the same year Nunnally (1955a) described an hypothetical matrix where sample heterogeneity with respect to persons resulted in very low correlations between variables while the obverse type correlations between the individuals were very high. Nunnally implied that ipsative scores are particularly valuable in yielding *Q*-type correlations which could reveal trends not apparent from *R*-type analyses. The way in which this matter depends upon the homogeneity of samples and the way in which it may be related to type of scale were not made explicit, however.

The growing interest in *Q* procedure has generated several methodological studies. Cohen (1957) has prepared a monograph which permits the investigator to read correlation coefficients between *Q* sorts, and Creaser (1955) has recommended a way for determining the amount an item should be weighted with respect to a given factor.

Goodling and Guthrie (1956) point out that the sample of items for *Q* sort should be selected in such a way as to provide maximum intersubject variability and minimum intrasubject variability. The question of intrasubject variability is one aspect of the reliability question, and this has been attacked directly by some investigators. For example, Hilden (1958) describes a sampling experiment where he begins with a universe of 1,575 statements from which he has randomly drawn 20 samples of 50 statements each. He had four graduate students provide self-ideal sorts for each of the 20 random sets and for the total population as well. The various scores, e.g., self-ideal, from any one set were correlated with each other, and the respective correlations were determined for the population. When the correlations for the random sets were compared with the correlations for the parent population, no reliable differences were found. From this one might infer that when using items such as these, a sample of 50 statements may be sufficient for *Q*-sort purposes.

There appears to be a general tendency among investigators to require their subjects to distribute their *Q* sorts in a quasi-normal fashion. This is in spite of the fact that Stephenson had recommended a flattened, bell-shaped distribution and that subsequent investigators had questioned the desirability of quasi-normal distributions. Jones (1956), for example, had noted that the free sorts of various groups differed appreciably from each other and that no group selected a bell-shaped distribution. Livson and Nichols (1956) had examined this problem from the standpoint of the number of discriminations that various shaped distributions involve, and noted that the more discriminations required, the greater the test-retest reliability of the sort. On the basis of this finding, these authors recommend that the *Q*-sort distribution should be rectangular. The issue of forced vs. unforced sorts has been discussed in numerous contexts, and no final agreement seems to have been reached. For example, Jones points out that there is no one preferred distribution, and Block (1956) believes, on the basis of his comparisons, that the forced sort method is equal or superior to free sorts.

Whether *Q* methodology will, as Stephenson proposed, create a psychology of the individual remains to be seen. From the standpoint of psychometry with its emphasis on individual differences or from the standpoint of psychoanalysis with its avoidance of formal instrumentation, *Q* methodology and the devices it includes do not provide an orthodox approach to the study of the individual. Certainly those particular psychologists who profess to be interested primarily in the individual have not rushed to apply this method to material which is still handled on an anecdotal or case history basis. Nevertheless, *Q* method's primary contributions to psychology appear to be in the study of psychotherapy and the related study of persons with personality disorders, and there are indications that this methodological emphasis can contribute to a broad study of personality and numerous related social problems. The growing acceptance of this methodological emphasis again reminds us that psychologists require flexible methods for their researches and will not wait for any orthodoxy.

REFERENCES

BECK, S. J. The six schizophrenias. *Res. Monogr. Amer. Orthopsychiat. Ass.*, 1954, No. 6.

BENDIG, A. W., & HAMLIN, R. M. The psychiatric validity of an inverted factor analysis of Rorschach scoring categories. *J. consult. Psychol.*, 1955, **19**, 183–188.

BLOCK, J. The difference between *Q* and *R*. *Psychol. Rev.*, 1955, **62**, 356–358.

BLOCK, J. A comparison of the forced and unforced *Q*-sorting procedures. *Educ. psychol. Measmt.*, 1956, **16**, 481–493.

BLOCK, J. A comparison between ipsative and normative ratings of personality. *J. abnorm. soc. Psychol.*, 1957, **54**, 50–54.

BLOCK, J., & PETERSON, P. Some personality correlates of confidence, caution, and speed in a decision situation. *J. abnorm. soc. Psychol.*, 1955, **51**, 34–41.

BROEN, W. E., JR. A factor-analytic study of religious attitudes. *J. abnorm. soc. Psychol.*, 1957, **54**, 176–179.

BUTLER, J. M., & HAIGH, G. V. Changes in the relation between self-concepts and ideal concepts. In C. R. Rogers & Rosalind F. Dymond (Eds.), *Psychotherapy and personality change*. Chicago: Univer. Chicago Press, 1954. Pp. 55–75.

CAPLAN, S. W. The effect of group counseling on junior high school boys' concepts of themselves in school. *J. counsel. Psychol.*, 1957, **4**, 124–128.

CARTWRIGHT, ROSALIND D. Effects of psychotherapy on self-consistency. *J. counsel. Psychol.*, 1957, **4**, 15–22.

CATTELL, R. B. Psychological measurement: Ipsative, normative, and interactive. *Psychol. Rev.*, 1944, **51**, 292–303.

CHASE, P. H. Self concepts in adjusted and maladjusted hospital patients. *J. consult. Psychol.*, 1957, **21**, 495–497.

COHEN, J. An aid in the computation of correlations based on *Q* sorts. *Psychol. Bull.*, 1957, **54**, 138–139.

CONGER, J. J., SAWREY, W. L., & KRAUSE, L. F. A reanalysis of Beck's "six schizophrenias." *J. consult. Psychol.*, 1956, **20**, 83–87.

CORSINI, R. J. Understanding and similarity in marriage. *J. abnorm. soc. Psychol.*, 1956, **52**, 327–332.

CREASER, J. W. An aid in calculating *Q*-sort factor-arrays. *J. clin. Psychol.*, 1955, **11**, 195–196.

DOLEYS, E. J., & KREGARMAN, J. Construct validity of the Chicago *Q*-sort: Frustration tolerance. *J. clin. Psychol.*, 1959, **15**, 177–179.

EDWARDS, A. L. Social desirability and *Q* sorts. *J. consult. Psychol.*, 1955, **19**, 462.

ENGEL, MARY. The stability of the self-concept in adolescence. *J. abnorm. soc. Psychol.*, 1959, **58**, 211–215.

EPSTEIN, S., & SMITH, R. Repression and insight as related to reaction to cartoons. *J. consult. Psychol.*, 1956, **20**, 391–395.

FAGAN, J., & GUTHRIE, G. M. Perception of self and of normality in schizophrenics. *J. clin. Psychol.*, 1959, **15**, 203–207.

FIEDLER, F. E. Factor analyses of psychoanalytic, nondirective, and Adlerian therapeutic relationships. *J. consult. Psychol.*, 1951, **15**, 32–38.

FISK, D. W., & VAN BUSKIRK, C. The stability of interpretations of sentence completion tests. *J. consult. Psychol.*, 1959, **23**, 177–180.

FRIEDMAN, I. Phenomenal, ideal, and projected conceptions of self. *J. abnorm. soc. soc. Psychol.*, 1955, **51**, 611–615.

GOODLING, R. A., & GUTHRIE, G. M. Some practical considerations in *Q*-sort item selection. *J. counsel. Psychol.*, 1956, **3**, 70–72.

GUILFORD, J. P. *Psychometric methods*. New York: McGraw-Hill, 1954.

HELFAND, I. Role taking in schizophrenia. *J. consult. Psychol.*, 1956, **20**, 37–41.

HILDEN, A. H. *Q*-sort correlation: Stability and random choice of statements. *J. consult. Psychol.*, 1958, **22**, 45–50.

JONES, A. Distributions of traits in current *Q*-sort methodology. *J. abnorm. soc. Psychol.*, 1956, **53**, 90–95.

KALIS, BETTY L., & BENNETT, LILLIAN F. The assessment of communication: The relation of clinical improvement to measured changes in communicative behavior. *J. consult. Psychol.*, 1957, **21**, 10–14.

KERLINGER, F. N. The attitude structure of the individual: A *Q*-study of the educational attitudes of professors and laymen. *Genet. psychol. Monogr.*, 1956, **53**, 283–329.

KOGAN, W. S., QUINN, R., AX, A. F., & RIPLEY, H. S. Some methodological problems in the quantification of clinical assessment by *Q* array. *J. consult. Psychol.*, 1957, **21**, 57–62.

LEVY, L. H. The meaning and generality of perceived actual-ideal discrepancies. *J. consult. Psychol.*, 1956, **20**, 396–398.

LIVSON, N. H., & NICHOLS, T. F. Discrimination and reliability in *Q*-sort personality descriptions. *J. abnorm. soc. Psychol.*, 1956, **52**, 159–165.

MORSH, J. E. The *Q*-sort technique as a group measure. *Educ. psychol. Measmt*, 1955, **15**, 390–395.

NAHINSKY, I. D. The relationship between the self-concept and the ideal-self concept as a measure of adjustment. *J. clin. Psychol.*, 1958, **14**, 360–364.

NUNNALLY, J. C. Some uses for "transpose" factor design in assessment research. *Educ. psychol. Measmt*, 1955, **15**, 240–245. (a)

NUNNALLY, J. C. A systematic approach to the construction of hypotheses about the process of psychotherapy. *J. consult. Psychol.*, 1955, **19**, 17–20. (b)

PETERSON, A. O. D., SNYDER, W. U., GUTHRIE, G. M., & RAY, W. S. Therapist factors: An exploratory investigation of therapeutic biases. *J. counsel. Psychol.*, 1958, **5**, 169–173.

RAWN, M. L. An experimental study of transference and resistance phenomena in psychoanalytically oriented psychotherapy. *J. clin. Psychol.*, 1958, **14**, 418–425.

REVIE, V. A. The effect of psychological case work on the teacher's concept of the pupil. *J. counsel. Psychol.*, 1956, **3**, 125–129.

REZNIKOFF, M., & TOOMEY, LAURA C. The weighted *Q* sort: A procedure for quantitatively estimating emotional disturbance and personality change. *J. consult. Psychol.*, 1958, **22**, 187–190.

ROGERS, A. H. The self-concept in paranoid schizophrenia. *J. clin. Psychol.*, 1958, **14**, 365–366.

ROGERS, C. R., & DYMOND, ROSALIND F. *Psychotherapy and personality change.* Chicago: Univer. Chicago Press, 1954.

SHEPHERD, IRMA L., & GUTHRIE, G. M. Attitudes of mothers of schizophrenic patients. *J. clin. Psychol.*, 1959, **15**, 212–215.

SHONTZ, F. C. Evaluative conceptualizations as the basis for clinical judgments. *J. consult. Psychol.*, 1956, **20**, 223–226.

STEPHENSON, W. *The study of behavior.* Chicago: Univer. Chicago Press, 1953.

STEWART, L. H. Manifest anxiety and mother-son identification. *J. clin. Psychol.*, 1958, **14**, 382–384.

SWEETLAND, A., & FRANK, G. A study of ideal psychological adjustment. *J. clin. Psychol.*, 1955, **11**, 391–394.

TAYLOR, D. M. Changes in the self concept without psychotherapy. *J. consult. Psychol.*, 1955, **19**, 205–209.

THRUSH, R. S. An agency in transition: The case study of a counseling center. *J. counsel. Psychol.*, 1957, **4**, 183–190.

TURNER, R. H., & VANDERLIPPE, R. H. Self-ideal congruence as an index of adjustment. *J. abnorm. soc. Psychol.*, 1958, **57**, 202–206.

WHITING, J. F. Needs, values, perceptions, and the nurse-patient relationship. *J. clin. Psychol.*, 1959, **15**, 146–150.

(Received January 4, 1960)

Psychological Bulletin
1961, Vol. 58, No. 2, 143–159

PSYCHOTHERAPY AS A LEARNING PROCESS

ALBERT BANDURA
Stanford University

While it is customary to conceptualize psychotherapy as a learning process, few therapists accept the full implications of this position. Indeed, this is best illustrated by the writings of the learning theorists themselves. Most of our current methods of psychotherapy represent an accumulation of more or less uncontrolled clinical experiences and, in many instances, those who have written about psychotherapy in terms of learning theory have merely substituted a new language; the practice remains essentially unchanged (Dollard, Auld, & White, 1954; Dollard & Miller, 1950; Shoben, 1949).

If one seriously subscribes to the view that psychotherapy is a learning process, the methods of treatment should be derived from our knowledge of learning and motivation. Such an orientation is likely to yield new techniques of treatment which, in many respects, may differ markedly from the procedures currently in use.

Psychotherapy rests on a very simple but fundamental assumption, i.e., human behavior is modifiable through psychological procedures. When skeptics raise the question, "Does psychotherapy work?" they may be responding in part to the mysticism that has come to surround the term. Perhaps the more meaningful question, and one which avoids the surplus meanings associated with the term "psychotherapy," is as follows: Can human behavior be modified through psychological means and if so, what are the learning mechanisms that mediate behavior change?

In the sections that follow, some of these learning mechanisms will be discussed, and studies in which systematic attempts have been made to apply these principles of learning to the area of psychotherapy will be reviewed. Since learning theory itself is still somewhat incomplete, the list of psychological processes by which changes in behavior can occur should not be regarded as exhaustive, nor are they necessarily without overlap.

COUNTERCONDITIONING

Of the various treatment methods derived from learning theory, those based on the principle of counterconditioning have been elaborated in greatest detail. Wolpe (1954, 1958, 1959) gives a thorough account of this method, and additional examples of cases treated in this manner are provided by Jones (1956), Lazarus and Rachman (1957), Meyer (1957), and Rachman (1959). Briefly, the principle involved is as follows: if strong responses which are incompatible with anxiety reactions can be made to occur in the presence of anxiety evoking cues, the incompatible responses will become attached to these cues and thereby weaken or eliminate the anxiety responses.

The first systematic psychotherapeutic application of this method was reported by Jones (1924b) in the treatment of Peter, a boy who showed severe phobic reactions to animals, fur objects, cotton, hair, and mechanical toys. Counterconditioning was achieved by feeding the child in the presence of initially small but gradually increasing anxiety-arousing stimuli. A rabbit in a cage was placed in the room at some distance so as not to disturb the boy's eating. Each day the rabbit was brought nearer to the table and eventually removed from the cage. During the final stage of treatment, the rabbit was placed on the feeding table and even in Peter's lap. Tests of generalization revealed that the fear responses had been effectively eliminated, not only toward the rabbit, but toward the previously feared furry objects as well.

In this connection, it would be interesting to speculate on the diagnosis and treatment Peter would have received had he been seen by Melanie Klein (1949) rather than by Mary Cover Jones!

It is interesting to note that while both Shoben (1949) and Wolpe (1958) propose a therapy based on the principle of counterconditioning, their treatment methods are radically different. According to Shoben, the patient discusses and thinks about stimulus situations that are anxiety provoking in the context of an interpersonal situation which simultaneously elicits positive affective responses from the patient. The therapeutic process consists in connecting the anxiety provoking stimuli, which are symbolically reproduced, with the comfort reaction made to the therapeutic relationship.

Shoben's paper represents primarily a counterconditioning interpretation of the behavior changes brought about through conventional forms of psychotherapy since, apart from highlighting the role of positive emotional reactions in the treatment process, no new techniques deliberately designed to facilitate relearning through counterconditioning are proposed.

This is not the case with Wolpe, who has made a radical departure from tradition. In his treatment, which he calls reciprocal inhibition, Wolpe makes systematic use of three types of responses which are antagonistic to, and therefore inhibitory of, anxiety. These are: assertive or approach responses, sexual responses, and relaxation responses.

On the basis of historical information, interview data, and psychological test responses, the therapist constructs an anxiety hierarchy, a ranked list of stimuli to which the patient reacts with anxiety. In the case of desensitization based on relaxation, the patient is hypnotized and given relaxation suggestions. He is then asked to imagine a scene representing the weakest item on the anxiety hierarchy and, if the relaxation is unimpaired, this is followed by having the patient imagine the next item on the list, and so on. Thus, the anxiety cues are gradually increased from session to session until the last phobic stimulus can be presented without impairing the relaxed state. Through this procedure, relaxation responses eventually come to be attached to the anxiety evoking stimuli.

Wolpe reports remarkable therapeutic success with a wide range of neurotic reactions treated on this counterconditioning principle. He also contends that the favorable outcomes achieved by the more conventional psychotherapeutic methods may result from the reciprocal inhibition of anxiety by strong positive responses evoked in the patient-therapist relationship.

Although the counterconditioning method has been employed most extensively in eliminating anxiety-motivated avoidance reactions and inhibitions, it has been used with some success in reducing maladaptive approach responses as well. In the latter case, the goal object is repeatedly associated with some form of aversive stimulus.

Raymond (1956), for example, used nausea as the aversion experience in the treatment of a patient who presented a fetish for handbags and perambulators which brought him into frequent contact with the law in that he repeatedly smeared mucus on ladies' handbags and destroyed perambulators by running into them with his motorcycle. Though the patient had undergone psychoanalytic treatment, and was fully aware of the origin and the sexual significance of his behavior, nevertheless, the fetish persisted.

The treatment consisted of showing the patient a collection of handbags, perambulators, and colored illustrations just before the onset of nausea produced by injections of apomorphine. The conditioning was repeated every 2 hours day and night for 1 week plus additional sessions 8 days and 6 months later.

Raymond reports that, not only was the fetish successfully eliminated, but also the patient showed a vast improvement in his social (and legal) relationships, was promoted to a more responsible position in his work, and no longer required the fetish fantasies to enable him to have sexual intercourse.

Nauseant drugs, especially emetine, have also been utilized as the unconditioned stimulus in the aversion treatment of alcoholism (Thirmann, 1949; Thompson & Bielinski, 1953; Voegtlen, 1940; Wallace, 1949). Usually 8 to 10 treatments in which the sight, smell, and taste of alcohol is associated with the onset of nausea is sufficient to produce abstinence. Of 1,000 or more cases on whom adequate follow-up data are reported, approximately 60% of the patients have been totally abstinent following the treatment. Voegtlen (1940) suggests that a few preventive treatments given at an interval of about 6 months may further improve the results yielded by this method.

Despite these encouraging findings, most psychotherapists are unlikely to be impressed since, in their opinion, the underlying causes for the alcoholism have in no way been modified by the conditioning procedure and, if anything, the mere removal of the alcoholism would tend to produce symptom substitution or other adverse effects. A full discussion of this issue will be presented later. In this particular context, however, several aspects of the Thompson and Bielinski (1953) data are worth noting. Among the alcoholic patients whom they treated, six "suffered from mental disorders not due to alcohol or associated deficiency states." It was planned, by the authors, to follow up the aversion treatment with psychotherapy for the underlying psychosis. This, however, proved unnecessary since all but one of the patients, a case of chronic mental deterioration, showed marked improvement and were in a state of remission.

Max (1935) employed a strong electric shock as the aversive stimulus in treating a patient who tended to display homosexual behavior following exposure to a fetishistic stimulus. Both the fetish and the homosexual behavior were removed through a series of avoidance conditioning sessions in which the patient was administered shock in the presence of the fetishistic object.

Wolpe (1958) has also reported favorable results with a similar procedure in the treatment of obsessions.

A further variation of the counterconditioning procedure has been developed by Mowrer and Mowrer

(1938) for use with enuretic patients. The device consists of a wired bed pad which sets off a loud buzzer and awakens the child as soon as micturition begins. Bladder tension thus becomes a cue for waking up which, in turn, is followed by sphincter contraction. Once bladder pressure becomes a stimulus for the more remote sphincter control response, the child is able to remain dry for relatively long periods of time without wakening.

Mowrer and Mowrer (1938) report complete success with 30 children treated by this method; similarly, Davidson and Douglass (1950) achieved highly successful results with 20 chronic enuretic children (15 cured, 5 markedly improved); of 5 cases treated by Morgan and Witmer (1939), 4 of the children not only gained full sphincter control, but also made a significant improvement in their social behavior. The one child with whom the conditioning approach had failed was later found to have bladder difficulties which required medical attention.

Some additional evidence for the efficacy of this method is provided by Martin and Kubly (1955) who obtained follow-up information from 118 of 220 parents who had treated their children at home with this type of conditioning apparatus. In 74% of the cases, according to the parents' replies, the treatment was successful.

Extinction

"When a learned response is repeated without reinforcement the strength of the tendency to perform that response undergoes a progressive decrease" (Dollard & Miller, 1950). Extinction involves the development of inhibitory potential which is composed of two components. The evocation of any reaction generates reactive inhibition (I_r) which presumably dissipates with time. When reactive inhibition (fatigue, etc.) reaches a high point, the cessation of activity alleviates this negative motivational state and any stimuli associated with the cessation of the response become conditioned inhibitors ($_sI_r$).

One factor that has been shown to influence the rate of extinction of maladaptive and anxiety-motivated behavior is the interval between extinction trials. In general, there tends to be little diminution in the strength of fear-motivated behavior when extinction trials are widely distributed, whereas under massed trials, reactive inhibition builds up rapidly and consequently extinction is accelerated (Calvin, Clifford, Clifford, Bolden, & Harvey, 1956; Edmonson & Amsel, 1954).

An illustration of the application of this principle is provided by Yates (1958) in the treatment of tics. Yates demonstrated, in line with the findings from laboratory studies of extinction under massed and distributed practice, that massed sessions in which the patient performed tics voluntarily followed by prolonged rest to allow for the dissipation of reactive inhibition was the most effective procedure for extinguishing the tics.

It should be noted that the extinction procedure employed by Yates is very similar to Dunlap's method of negative practice, in which the subject reproduces the negative behaviors voluntarily without reinforcement (Dunlap, 1932; Lehner, 1954). This method has been applied most frequently, with varying degrees of success, to the treatment of speech disorders (Fishman, 1937; Meissner, 1946; Rutherford, 1940; Sheehan, 1951; Sheehan & Voas, 1957). If the effectiveness of this psychotherapeutic technique is due primarily to extinction, as suggested by Yates' study, the usual practice of terminating a treatment session before the subject becomes fatigued (Lehner, 1954), would have the effect of reducing the rate of extinction, and may in part account for the divergent results yielded by this method.

Additional examples of the therapeutic application of extinction procedures are provided by Jones (1955), and most recently by C. D. Williams (1959).

Most of the conventional forms of psychotherapy rely heavily on extinction effects although the therapist may not label these as such. For example, many therapists consider *permissiveness* to be a necessary condition of therapeutic change (Alexander, 1956; Dollard & Miller, 1950; Rogers, 1951). It is expected that when a patient expresses thoughts or feelings that provoke anxiety or guilt and the therapist does not disapprove, criticize, or withdraw interest, the fear or guilt will be gradually weakened or extinguished. The extinction effects are believed to generalize to thoughts concerning related topics that were originally inhibited, and to verbal and physical forms of behavior as well (Dollard & Miller, 1950).

Some evidence for the relationship between permissiveness and the extinction of anxiety is provided in two studies recently reported by Dittes (1957a, 1957b). In one study (1957b) involving an analysis of patient-therapist interaction sequences, Dittes found that permissive responses on the part of the therapist were followed by a corresponding decrease in the patient's anxiety (as measured by the GSR) and the occurrence of avoidance behaviors. A sequential analysis of the therapeutic sessions (Dittes, 1957a), revealed that, at the onset of treatment, sex expressions were accompanied by strong anxiety reactions; under the cumulative effects of permissiveness, the anxiety gradually extinguished.

In contrast to counterconditioning, extinction is likely to be a less effective and a more time consuming method for eliminating maladaptive behavior (Jones, 1924a; Dollard & Miller, 1950); in the case of conventional interview therapy, the relatively long intervals between interview sessions, and the ritualistic adherence to the 50-minute hour may further reduce the occurrence of extinction effects.

Discrimination Learning

Human functioning would be extremely difficult and inefficient if a person had to learn appropriate behavior for every specific situation he encountered. Fortunately, patterns of behavior learned in one situation will transfer or generalize to other similar situations. On the other hand, if a person overgeneralizes from one situation to another, or if the generalization is based on superficial or irrelevant cues, behavior becomes inappropriate and maladaptive.

In most theories of psychotherapy, therefore, discrimination learning, believed to be accomplished through the gaining of awareness or insight, receives emphasis (Dollard & Miller, 1950; Fenichel, 1941; Rogers, 1951; Sullivan, 1953). It is generally assumed that if a patient is aware of the cues producing his behavior, of the responses he is making, and of the reasons that he responds the way he does, his behavior will become more susceptible to verbally-mediated control. Voluntarily guided, discriminative behavior will replace the automatic, overgeneralized reactions.

While this view is widely accepted, as evidenced in the almost exclusive reliance on interview procedures and on interpretative or labeling techniques, a few therapists (Alexander & French, 1946) have questioned the importance attached to awareness in producing modifications in behavior. Whereas most psychoanalysts (Fenichel, 1941), as well as therapists representing other points of view (Fromm-Reichmann, 1950; Sullivan, 1953) consider insight a precondition of behavior change, Alexander and French consider insight or awareness a result of change rather than its cause. That is, as the patient's anxieties are gradually reduced through the permissive conditions of treatment, formerly inhibited thoughts are gradually restored to awareness.

Evidence obtained through controlled laboratory studies concerning the value of awareness in increasing the precision of discrimination has so far been largely negative or at least equivocal (Adams, 1957; Erikson, 1958; Razran, 1949). A study by Lacy and Smith (1954), in which they found aware subjects generalized anxiety reactions less extensively than did subjects who were unaware of the conditioned stimulus provides evidence that awareness may aid discrimination. However, other aspects of their findings (e.g., the magnitude of the anxiety reactions to the generalization stimuli were greater than they were to the conditioned stimulus itself) indicate the need for replication.

If future research continues to demonstrate that awareness exerts little influence on the acquisition, generalization, and modification of behavior, such negative results would cast serious doubt on the value of currently popular psychotherapeutic procedures whose primary aim is the development of insight.

Methods of Reward

Most theories of psychotherapy are based on the assumption that the patient has a repertoire of previously learned positive habits available to him, but that these adaptive patterns are inhibited or blocked by competing responses motivated by anxiety or guilt. The goal of therapy, then, is to reduce the severity of the internal inhibitory controls, thus allowing the healthy patterns of behavior to emerge. Hence, the role of the therapist is to create permissive conditions under which the patient's "normal growth potentialities" are set free (Rogers, 1951). The fact that most of our theories of personality and therapeutic procedures have been developed primarily through work with oversocialized, neurotic patients may account in part for the prevalence of this view.

There is a large class of disorders (the undersocialized, antisocial personalities whose behavior reflects a failure of the socialization process) for whom this model of personality and accompanying techniques of treatment are quite inappropriate (Bandura & Walters, 1959; Schmideberg, 1959). Such antisocial personalities are likely to present *learning deficits*, consequently the goal of therapy is the acquisition of secondary motives and the development of internal restraint habits. That antisocial patients prove unresponsive to psychotherapeutic methods developed for the treatment of oversocialized neurotics has been demonstrated in a number of studies comparing patients who remain in treatment with those who terminate treatment prematurely (Rubenstein & Lorr, 1956). It is for this class of patients that the greatest departures from traditional treatment methods is needed.

While counterconditioning, extinction, and discrimination learning may be effective ways of removing neurotic inhibitions, these methods may be of relatively little value in developing new positive habits. Primary and secondary rewards in the form of the therapist's interest and approval may play an important, if not indispensable, role in the treatment process. Once the patient has learned to want the interest and approval of the therapist, these rewards may then be used to promote the acquisition of new patterns of behavior. For certain classes of patients such as schizophrenics (Atkinson, 1957; Peters, 1953; Robinson, 1957) and delinquents (Cairns, 1959), who are either unresponsive to, or fearful of, social rewards, the therapist may have to rely initially on primary rewards in the treatment process.

An ingenious study by Peters and Jenkins (1954) illustrates the application of this principle in the treatment of schizophrenic patients. Chronic patients from closed wards were administered subshock injections of insulin designed to induce the hunger drive. The patients were then encouraged to solve a series of graded problem tasks with fudge as the reward. This program was followed 5 days a week for 3 months.

Initially the tasks involved simple mazes and obstruction problems in which the patients obtained the food reward directly upon successful completion of the problem. Tasks of gradually increasing difficulty were then administered involving multiple-choice learning and verbal-reasoning problems in which the experimenter personally mediated the primary rewards. After several weeks of such problem solving activities the insulin injections were discontinued and social rewards, which by this time had become more effective, were used in solving interpersonal problems that the patients were likely to encounter in their daily activities both inside and outside the hospital setting.

Comparison of the treated group with control groups, designed to isolate the effects of insulin and special attention, revealed that the patients in the reward group improved significantly in their social relationships in the hospital, whereas the patients in the control groups showed no such change.

King and Armitage (1958) report a somewhat similar study in which severely withdrawn schizophrenic patients were treated with operant conditioning methods; candy and cigarettes served as the primary rewards for eliciting and maintaining increasingly complex forms of behavior, i.e., psychomotor, verbal, and interpersonal responses. Unlike the Peters and Jenkins study, no attempt was made to manipulate the level of primary motivation.

An interesting feature of the experimental design was the inclusion of a group of patients who were treated with conventional interview therapy, as well as a recreational therapy and a no-therapy control group. It was found that the operant group, in relation to similar patients in the three control groups, made significantly more clinical improvement.

Skinner (1956b) and Lindsley (1956) working with adult psychotics, and Ferster (1959) working with autistic children, have been successful in developing substantial amounts of reality-oriented behavior in their patients through the use of reward. So far their work has been concerned primarily with the effect of schedules of reinforcement on the rate of evocation of simple impersonal reactions. There is every indication, however, that by varying the contingency of the reward (e.g., the patient must respond in certain specified ways to the behavior of another individual in order to produce the reward) adaptive interpersonal behaviors can be developed as well (Azran & Lindsley, 1956).

The effectiveness of social reinforcers in modifying behavior has been demonstrated repeatedly in verbal conditioning experiments (Krasner, 1958; Salzinger, 1959). Encouraged by these findings, several therapists have begun to experiment with operant conditioning as a method of treatment in its own right (Tilton, 1956; Ullman, Krasner, & Collins, in press; R. I. Williams, 1959); the operant conditioning studies cited earlier are also illustrative of this trend.

So far the study of generalization and permanence of behavior changes brought about through operant conditioning methods has received relatively little attention and the scanty data available are equivocal (Rogers, 1960; Sarason, 1957; Weide, 1959). The lack of consistency in results is hardly surprising considering that the experimental manipulations in many of the conditioning studies are barely sufficient to demonstrate conditioning effects, let alone generalization of changes to new situations. On the other hand, investigators who have conducted more intensive reinforcement sessions, in an effort to test the efficacy of operant conditioning methods as a therapeutic technique, have found significant changes in patients' interpersonal behavior in extra-experimental situations (King & Armitage, 1958; Peters & Jenkins, 1954; Ullman et al., in press). These findings are particularly noteworthy since the response classes involved are similar to those psychotherapists are primarily concerned in modifying through interview forms of treatment. If the favorable results yielded by these studies are replicated in future investigations, it is likely that the next few years will witness an increasing reliance on conditioning forms of psychotherapy, particularly in the treatment of psychotic patients.

At this point it might also be noted that, consistent with the results from verbal conditioning experiments, content analyses of psychotherapeutic interviews (Bandura, Lipsher, & Miller, 1960; Murray, 1956) suggest that many of the changes observed in psychotherapy, at least insofar as the patients' verbal behavior is concerned, can be accounted for in terms of the therapists' direct, although usually unwitting, reward and punishment of the patients' expressions.

Punishment

While positive habits can be readily developed through reward, the elimination of socially disapproved habits, which becomes very much an issue in the treatment of antisocial personalities, poses a far more complex problem.

The elimination of socially disapproved behaviors can be accomplished in several ways. They may be consistently unrewarded and thus extinguished. However, antisocial behavior, particularly of an extreme form, cannot simply be ignored in the hope that it will gradually extinguish. Furthermore, since the successful execution of antisocial acts may bring substantial material rewards as well as the approval and admiration of associates, it is extremely unlikely that such behavior would ever extinguish.

Although punishment may lead to the rapid disappearance of socially disapproved behavior, its effects are far more complex (Estes, 1944; Solomon, Kamin, & Wynne, 1953). If a person is punished for some socially disapproved habit, the impulse to perform the act becomes, through its association with punishment, a stimulus for anxiety. This anxiety then motivates competing responses which, if sufficiently strong, prevent the occurrence of, or inhibit, the disapproved behavior. Inhibited responses may not, however, thereby lose their strength, and may reappear in situations where the threat of punishment is weaker. Punishment may, in fact, prevent the extinction of a habit; if a habit is completely inhibited, it cannot occur and therefore cannot go unrewarded.

Several other factors point to the futility of punishment as a means of correcting many antisocial patterns. The threat of punishment is very likely to elicit conformity; indeed, the patient may obligingly do whatever he is told to do in order to avoid immediate difficulties. This does not mean, however, that he has acquired a set of sanctions that will be of service to him once he is outside the treatment situation. In fact, rather than leading to the development of internal controls, such methods are likely only to increase the patient's reliance on external restraints. Moreover, under these conditions, the majority of patients will develop the attitude that they will do only what they are told to do—and then often only half-heartedly—and that they will do as they please once they are free from the therapist's supervision (Bandura & Walters, 1959).

In addition, punishment may serve only to intensify hostility and other negative motivations and thus may further instigate the antisocial person to display the very behaviors that the punishment was intended to bring under control.

Mild aversive stimuli have been utilized, of course, in the treatment of voluntary patients who express a desire to rid themselves of specific debilitating conditions.

Liversedge and Sylvester (1955), for example, successfully treated seven cases of writer's cramp by means of a retraining procedure involving electric shock. In order to remove tremors, one component of the motor disorder, the patients were required to insert a stylus into a series of progressively smaller holes; each time the stylus made contact with the side of the hole the patients received a mild shock. The removal of the spasm component of the disorder was obtained in two ways. First, the patients traced various line patterns (similar to the movements required in writing) on a metal plate with a stylus, and any deviation from the path produced a shock. Following training on the apparatus, the subjects then wrote with an electrified pen which delivered a shock whenever excessive thumb pressure was applied.

Liversedge and Sylvester report that following the retraining the patients were able to resume work; a follow-up several months later indicated that the improvement was being maintained.

The aversive forms of therapy, described earlier in the section on counterconditioning procedures, also make use of mild punishment.

Social Imitation

Although a certain amount of learning takes place through direct training and reward, a good deal of a person's behavior repertoire may be acquired through imitation of what he observes in others. If this is the case, social imitation may serve as an effective vehicle for the transmission of prosocial behavior patterns in the treatment of antisocial patients.

Merely providing a model for imitation is not, however, sufficient. Even though the therapist exhibits the kinds of behaviors that he wants the patient to learn, this is likely to have little influence on him if he rejects the therapist as a model. Affectional nurturance is believed to be an important precondition for imitative learning to occur, in that affectional rewards increase the secondary reinforcing properties of the model, and thus predispose the imitator to pattern his behavior after the rewarding person (Mowrer, 1950; Sears, 1957; Whiting, 1954). Some positive evidence for the influence of social rewards on imitation is provided by Bandura and Huston (in press) in a recent study of identification as a process of incidental imitation.

In this investigation preschool children performed an orienting task but, unlike most incidental learning studies, the experimenter performed the diverting task as well, and the extent to which the subjects patterned their behavior after that of the experimenter-model was measured.

A two-choice discrimination problem similar to the one employed by Miller and Dollard (1941) in their experiments of social imitation was used as the diverting task. On each trial, one of two boxes was loaded with two rewards (small multicolor pictures of animals) and the object of the game was to guess which box contained the stickers. The experimenter-model (*M*) always had her turn first and in each instance chose the reward box. During *M*'s trial, the subject remained at the starting point where he could observe the *M*'s behavior. On each discrimination trial *M* exhibited certain verbal, motor, and aggressive patterns of behavior that were totally irrelevant to the task to which the subject's attention was directed. At the starting point, for example, *M* made a verbal response and then marched slowly toward the box containing the stickers, repeating, "March, march, march." On the lid of each box was a rubber doll which *M* knocked off aggressively when she reached the designated box. She then paused briefly, remarked, "Open the box," removed one sticker, and pasted it on a pastoral scene which hung on the wall immediately behind the boxes. The subject then took his turn and the number of *M*'s behaviors performed by the subject was recorded.

A control group was included in order to, (*a*) provide a check on whether the subjects' performances reflected genuine imitative learning or merely the chance occurrence of behaviors high in the subjects' response hierarchies, and (*b*) to determine whether subjects would adopt certain aspects of *M*'s behavior which involved considerable delay in reward. With the controls, therefore, *M* walked to the box, choosing a highly circuitous route along the sides of the experimental room; instead of aggressing toward the doll, she lifted it gently off the container.

The results of this study indicate that, insofar as preschool children are concerened, a good deal of incidental imitation of the behaviors displayed by an adult model does occur. Of the subjects in the experimental group, 88% adopted the *M*'s aggressive behavior, 44% imitated the marching, and 28% reproduced *M*'s verbalizations. In contrast, none of the control subjects behaved aggressively, marched, or verbalized, while 75% of the controls imitated the circuitous route to the containers.

In order to test the hypothesis that children who experience a rewarding relationship with an adult model adopt more of the model's behavior than do children who experience a relatively distant and cold relationship, half the subjects in the experiment were assigned to a nurturant condition; the other half of the subjects to a nonnurturant condition.

During the nurturant sessions, which preceded the incidental learning, *M* played with subject, she responded readily to the subject's bids for attention, and in other ways fostered a consistently warm and rewarding interaction with the child. In contrast, during the nonnurturant sessions, the subject played alone while *M* busied herself with paperwork at a desk in the far corner of the room.

Consistent with the hypothesis, it was found that subjects who experienced the rewarding interaction with *M* adopted significantly more of *M*'s behavior than did subjects who were in the nonnurturance condition.

A more crucial test of the transmission of behavior patterns through the process of social imitation involves the delayed generalization of imitative responses to new situations in which the model is absent. A study of this type just completed, provides strong evidence that observation of the cues produced by the behavior of others is an effective means of eliciting responses for which the original probability is very low (Bandura, Ross, & Ross, in press).

Empirical studies of the correlates of strong and weak identification with parents, lend additional support to the theory that rewards promote imitative learning. Boys whose fathers are highly rewarding and affectionate have been found to adopt the father-role in doll-play activities (Sears, 1953), to show father-son similarity in response to items on a personality questionnaire (Payne & Mussen, 1956), and to display masculine behaviors (Mussen & Distler, 1956, 1960) to a greater extent than boys whose fathers are relatively cold and nonrewarding.

The treatment of older unsocialized delinquents is a difficult task, since they are relatively self-sufficient and do not readily seek involvement with a therapist. In many cases, socialization can be accomplished only through residental care and treatment. In the treatment home, the therapist can personally administer many of the primary rewards and mediate between the boys' needs and gratifications. Through the repeated association with rewarding experiences for the boy, many of the therapist's attitudes and actions will acquire secondary reward value, and thus the patient will be motivated to reproduce these attitudes and actions in himself. Once these attitudes and values have been thus accepted, the boy's inhibition of antisocial tendencies will function independently of the therapist.

While treatment through social imitation has been suggested as a method for modifying antisocial patterns, it can be an effective procedure for the treatment of other forms of disorders as well. Jones (1924a), for example, found that the social example of children reacting normally to stimuli feared by another child was effective, in some instances, in eliminating such phobic reactions. In fact, next to counterconditioning, the method of social imitation proved to be most effective in eliminating inappropriate fears.

There is some suggestive evidence that by providing high prestige models and thus increasing the reinforcement value of the imitatee's behavior, the effectiveness of this method in promoting favorable adjustive patterns of behavior may be further increased (Jones, 1924a; Mausner, 1953, 1954; Miller & Dollard, 1941).

During the course of conventional psychotherapy, the patient is exposed to many incidental cues involving the therapist's values, attitudes, and patterns of behavior. They are incidental only because they are usually considered secondary or irrelevant to the task of resolving the patient's problems. Nevertheless, some of the changes observed in the patient's behavior may result, not so much from the intentional interaction between the patient and the therapist, but rather from active learning by the patient of the therapist's attitudes and values which the therapist never directly attempted to transmit. This is partially corroborated by Rosenthal (1955) who found that, in spite of the usual precautions taken by therapists to avoid imposing their values on their clients, the patients who were judged as showing the greatest improvement changed their moral values (in the areas of sex, aggression, and authority) in the direction of the values of their therapists, whereas patients who were unimproved became less like the therapist in values.

Factors Impeding Integration

In reviewing the literature on psychotherapy, it becomes clearly evident that learning theory and general psychology have exerted a remarkably minor influence on the practice of psychotherapy and, apart from the recent interest in Skinner's operant conditioning methods (Krasner, 1955; Skinner, 1953), most of the recent serious attempts to apply learning principles to clinical practice have been made by European psychotherapists (Jones, 1956; Lazarus & Rachman, 1957; Liversedge & Sylvester, 1955; Meyer, 1957; Rachman, 1959; Raymond, 1956; Wolpe, 1958; Yates, 1958). This isolation of the methods of treatment from our knowledge of learning and motivation will continue to exist for some time since there are several prevalent attitudes that impede adequate integration.

In the first place, the deliberate use of the principles of learning in the modification of human behavior implies, for most psychotherapists, manipulation and control of the patient, and control is seen by them as antihumanistic and, therefore, bad. Thus, advocates of a learning approach to psychotherapy are often charged with treating human beings as though they were rats or pigeons and of leading on the road to Orwell's *1984*.

This does not mean that psychotherapists do not influence and control their patients' behavior. On the contrary. In any interpersonal interaction, and psychotherapy is no exception, people influence and control one another (Frank, 1959; Skinner, 1956a). Although the patient's control of the therapist has not as yet been studied (such control is evident when patients subtly reward the therapist with interesting historical material and thereby avoid the discussion of their current interpersonal problems), there is considerable evidence that the therapist exercises personal control over his patients. A brief examination of interview protocols of patients treated by therapists representing differing theoretical orientations, clearly reveals that the patients have been thoroughly conditioned in their therapists' idiosyncratic languages. Client-centered patients, for example, tend to produce the client-centered terminology, theory, and goals, and their interview content shows little or no overlap with that of patients seen in psychoanalysis who, in turn, tend to speak the language of psychoanalytic theory (Heine, 1950). Even more direct evidence of the therapists' controlling influence is provided in studies of patient-therapist interactions (Bandura et al., 1960; Murray, 1956; Rogers, 1960). The results of these studies show that the therapist not only controls the patient by rewarding him with interest and approval when the patient behaves in a fashion the therapist desires, but that he also controls through punishment, in the form of mild disapproval and withdrawal of interest, when the patient behaves in ways that are threatening to the therapist or run counter to his goals.

One difficulty in understanding the changes that occur in the course of psychotherapy is that the independent variable, i.e., the therapist's behavior, is often vaguely or only partially defined. In an effort to minimize or to deny the therapist's directive influence on the patient, the therapist is typically depicted as a "catalyst" who, in some mysterious way, sets free positive adjustive patterns of behavior or similar outcomes usually described in very general and highly socially desirable terms.

It has been suggested, in the material presented in the preceding sections, that many of the changes that occur in psychotherapy derive from the unwitting application of well-known principles of learning. However, the occurrence of the necessary conditions for learning is more by accident than by intent and, perhaps, a more deliberate application of our knowledge of the learning process to psychotherapy would yield far more effective results.

The predominant approach in the development of psychotherapeutic procedures has been the "school" approach. A similar trend is noted in the treatment methods being derived from learning theory. Wolpe, for example, has selected the principle of counterconditioning and built a "school" of psychotherapy around it; Dollard and Miller have focused on extinction and discrimination learning; and the followers of Skinner rely almost entirely on methods of reward. This stress on a few learning principles at the expense of neglecting other relevant ones will serve only to limit the effectiveness of psychotherapy.

A second factor that may account for the discontinuity between general psychology and psychotherapeutic practice is that the model of personality to which most therapists subscribe is somewhat dissonant with the currently developing principles of behavior.

In their formulations of personality functioning, psychotherapists are inclined to appeal to a variety of inner explanatory processes. In contrast, learning theorists view the organism as a far more mechanistic and simpler system, and consequently their formulations tend to be expressed for the most part in terms of antecedent-consequent relationships without reference to inner states.

> Symptoms are learned S-R connections; once they are extinguished or deconditioned treatment is complete. Such treatment is based exclusively on present factors; like Lewin's theory, this one is a-historical. Nonverbal methods are favored over verbal ones, although a minor place is reserved for verbal methods of extinction and reconditioning. Concern is with *function*, not with *content*. The main difference between the two theories arises over the question of "symptomatic" treatment. According to orthodox theory, this is useless unless the underlying complexes are attacked. According to the present theory, there is no evidence for these putative complexes, and symptomatic treatment is all that is required (Eysenck, 1957, pp. 267–268). (Quoted by permission of Frederick A. Praeger, Inc.)

Changes in behavior brought about through such methods as counterconditioning are apt to be viewed by the "dynamically-oriented" therapist, as being not only superficial, "symptomatic" treatment, in that the basic underlying instigators of the behavior remain unchanged, but also potentially dangerous, since the direct elimination of a symptom may precipitate more seriously disturbed behavior.

This expectation receives little support from the generally favorable outcomes reported in the studies reviewed in this paper. In most cases where follow-up data were available to assess the long-term effects of the therapy, the patients, many of whom had been treated by conventional methods with little benefit, had evidently become considerably more effective in their social, vocational, and psychosexual adjustment. On the whole the evidence, while open to error, suggests that no matter what the origin of the maladaptive behavior may be, a change in behavior brought about through learning procedures may be all that is necessary for the alleviation of most forms of emotional disorders.

As Mowrer (1950) very aptly points out, the "symptom-underlying cause" formulation may represent inappropriate medical analogizing. Whether or not a given behavior will be considered normal or a symtom of an underlying disturbance will depend on whether or not somebody objects to the behavior. For example, aggressiveness on the part of children may be encouraged and considered a sign of healthy development by the parents, while the same behavior is viewed by school authorities and society as a symptom of a personality disorder (Bandura & Walters, 1959). Furthermore, behavior considered to be normal at one stage in development may be reregarded as a "symptom of a personality disturbance" at a later period. In this connection it is very appropriate to repeat Mowrer's (1950) query: "And when does persisting behavior of this kind suddenly cease to be normal and become a symptom" (p. 474).

Thus, while a high fever is generally considered a sign of an underlying disease process regardless of when or where it occurs, whether a specific behavior will be viewed as normal or as a symptom of an underlying pathology is not independent of who makes the judgement, the social context in which the behavior occurs, the age of the person, as well as many other factors.

Another important difference between physical pathology and behavior pathology usually overlooked is that, in the case of most behavior disorders, it is not the underlying motivations that need to be altered or removed, but rather the ways in which the patient has learned to gratify his needs (Rotter, 1954). Thus, for example, if a patient displays deviant sexual behavior, the goal is not the removal of the underlying causes, i.e., sexual motivation, but rather the substitution of more socially approved instrumental and goal responses.

It might also be mentioned in passing, that, in the currently popular forms of psychotherapy, the role assumed by the therapist may bring him a good many direct or fantasied personal gratifications. In the course of treatment the patient may express considerable affection and admiration for the therapist, he may assign the therapist an omniscient status, and the reconstruction of the patient's history may be an intellectually stimulating activity. On the other hand, the methods derived from learning theory place the therapist in a less glamorous role, and this in itself may create some reluctance on the part of psychotherapists to part with the procedures currently in use.

Which of the two conceptual theories of personality—the psychodynamic or the social learning theory—is the more useful in generating effective procedures for the modification of human behavior remains to be demonstrated. While it is possible to present logical arguments and impressive clinical evidence for the efficiency of either approach, the best proving ground is the laboratory.

In evaluating psychotherapeutic methods, the common practice is to compare changes in a treated group with those of a nontreated control group. One drawback of this approach is that, while it answers the question as to whether or not a particular treatment is more effective than no intervention in producing changes along specific dimensions for certain classes of patients, it does not provide evidence concerning the relative effectiveness of alternative forms of psychotherapy.

It would be far more informative if, in future psychotherapy research, radically different forms of treatment were compared (King & Armitage, 1958; Rogers, 1959), since this approach would lead to a more rapid discarding of those of our cherished psychotherapeutic rituals that prove to be ineffective in, or even a handicap to, the successful treatment of emotional disorders.

REFERENCES

ADAMS, J. K. Laboratory studies of behavior without awareness. *Psychol. Bull.*, 1957, **54**, 393–405.

ALEXANDER, F. *Psychoanalysis and psychotherapy.* New York: Norton, 1956.

ALEXANDER, F., & FRENCH, M. T. *Psychoanalytic therapy.* New York: Ronald, 1946.

ATKINSON, RITA L. Paired-associate learning by schizophrenic and normal subjects under conditions of verbal reward and verbal punishment. Unpublished doctoral dissertation, Indiana University, 1957.

AZRAN, N. H., & LINDSLEY, O. R. The reinforcement of cooperation between children. *J. abnorm. soc. Psychol.*, 1956, **52**, 100–102.

BANDURA A., & HUSTON, ALETHA, C. Identification as a process of incidental learning. *J. abnorm. soc. Psychol.*, in press.

BANDURA, A., LIPSHER, D. H., & MILLER, PAULA, E. Psychotherapists' approach-avoidance reactions to patients' expressions of hostility. *J. consult. Psychol.*, 1960, **24**, 1–8.

BANDURA, A., ROSS, DOROTHEA, & ROSS, SHEILA, A. Transmission of aggression through imitation of aggressive models. *J. abnorm. soc. Psychol.*, in press.

BANDURA, A., & WALTERS, R. H. *Adolescent aggression.* New York: Ronald, 1959.

CAIRNS, R. B. The influence of dependency-anxiety on the effectiveness of social reinforcers. Unpublished doctoral dissertation, Stanford University, 1959.

CALVIN, A. D., CLIFFORD, L. T., CLIFFORD, B., BOLDEN, L., & HARVEY, J. Experimental validation of conditioned inhibition. *Psychol. Rep.*, 1956, **2**, 51–56.

DAVIDSON, J. R., & DOUGLASS, E. Nocturnal enuresis: A special approach to treatment. *British med. J.*, 1950, **1**, 1345–1347.

DITTES, J. E. Extinction during psychotherapy of GSR accompanying "embarrassing" statements. *J. abnorm. soc. Psychol.*, 1957, **54**, 187–191. (a)

DITTES, J. E. Galvanic skin responses as a measure of patient's reaction to therapist's permissiveness. *J. abnorm. soc. Psychol.*, 1957, **55**, 295–303. (b)

DOLLARD, J., AULD, F., & WHITE, A. M. *Steps in psychotherapy.* New York: Macmillan, 1954.

DOLLARD, J., & MILLER, N. E. *Personality and psychotherapy.* New York: McGraw-Hill, 1950.

DUNLAP, K. *Habits, their making and unmaking.* New York: Liveright, 1932.

EDMONSON, B. W., & AMSEL, A. The effects of massing and distribution of extinction trials on the persistence of a fear-motivated instrumental response. *J. comp. physiol. Psychol.*, 1954, **47**, 117–123.

ERIKSON, C. W. Unconscious processes. In M. R. Jones (Ed.), *Nebraska symposuim on motivation.* Lincoln: Univer. Nebraska Press, 1958.

ESTES, W. K. An experimental study of punishment. *Psychol. Monogr.*, 1944, **57** (3, Whole No. 363).

EYSENCK, H. J. *The dynamics of anxiety and hysteria.* New York: Praeger, 1957.

FENICHEL, O. *Problems of psychoanalytic technique.* (Trans. by D. Brunswick) New York: Psychoanalytic Quarterly, 1941.

FERSTER, C. B. Development of normal behavioral processes in autistic children. *Res. relat. Child.*, 1959, No. 9, 30. (Abstract)

FISHMAN, H. C. A study of the efficiency of negative practice as a corrective for stammering. *J. speech Dis.*, 1937, **2**, 67–72.

FRANK, J. D. The dynamics of the psychotherapeutic relationship. *Psychiatry*, 1959, **22**, 17–39.

FROMM-REICHMANN, FRIEDA. *Principle of intensive psychotherapy*. Chicago: Univer. Chicago Press, 1950.

HEINE, R. W. An investigation of the relationship between change in personality from psychotherapy as reported by patients and the factors seen by patients as producing change. Unpublished doctoral dissertation, University of Chicago, 1950.

JONES, E. L. Exploration of experimental extinction and spontaneous recovery in stuttering. In W. Johnson (Ed.), *Stuttering in children and adults*. Minneapolis: Univer. Minnesota Press, 1955.

JONES, H. G. The application of conditioning and learning techniques to the treatment of a psychiatric patient. *J. abnorm. soc. Psychol.*, 1956, **52**, 414–419.

JONES, MARY C. The elimination of childrens' fears. *J. exp. Psychol.*, 1924, **7**, 382–390. (a)

JONES, MARY C. A laboratory study of fear: The case of Peter. *J. genet. Psychol.*, 1924, **31**, 308–315. (b)

KING, G. F., & ARMITAGE, S. G. An operant-interpersonal therapeutic approach to schizophrenics of extreme pathology. *Amer. Psychologist*, 1958, **13**, 358. (Abstract)

KLEIN, MELANIE. *The psycho-analysis of children*. London: Hogarth, 1949.

KRASNER, L. The use of generalized reinforcers in psychotherapy research. *Psychol. Rep.*, 1955, **1**, 19–25.

KRASNER, L. Studies of the conditioning of verbal behavior. *Psychol. Bull.*, 1958, **55**, 148–170.

LACEY, J. I., & SMITH, R. I. Conditioning and generalization of unconscious anxiety. *Science*, 1954, **120**, 1–8.

LAZARUS, A. A., & RACHMAN, S. The use of systematic desensitization in psychotherapy. *S. Afr. med. J.*, 1957, **32**, 934–937.

LEHNER, G. F. J. Negative practice as a psychotherapeutic technique. *J. gen. Psychol.*, 1954, **51**, 69–82.

LINDSLEY, O. R. Operant conditioning methods applied to research in chronic schizophrenia. *Psychiat. res. Rep.*, 1956, **5**, 118–138.

LIVERSEDGE, L. A., & SYLVESTER, J. D. Conditioning techniques in the treatment of writer's cramp. *Lancet*, 1955, **1**, 1147–1149.

MARTIN, B., & KUBLY, DELORES. Results of treatment of enuresis by a conditioned response method. *J. consult. Psychol.*, 1955, **19**, 71–73.

MAUSNER, B. Studies in social interaction: III. The effect of variation in one partner's prestige on the interaction of observer pairs. *J. appl. Psychol.*, 1953, **37**, 391–393.

MAUSNER, B. The effect of one partner's success in a relevant task on the interaction of observer pairs. *J. abnorm. soc. Psychol.*, 1954, **49**, 557–560.

MAX, L. W. Breaking up a homosexual fixation by the conditioned reaction technique: A case study. *Psychol. Bull.*, 1935, **32**, 734.

MEISSNER, J. H. The relationship between voluntary nonfluency and stuttering. *J. speech Dis.*, 1946, **11**, 13–33.

MEYER, V. The treatment of two phobic patients on the basis of learning principles: Case report. *J. abnorm. soc. Psychol.*, 1957, **55**, 261–266.

MILLER, N. E., & DOLLARD, J. *Social learning and imitation*. New Haven: Yale Univer. Press, 1941.

MORGAN, J. J. B., & WITMER, F. J. The treatment of enuresis by the conditioned reaction technique. *J. genet. Psychol.*, 1939, **55**, 59–65.

MOWRER, O. H. *Learning theory and personality dynamics*. New York: Ronald, 1950.

MOWRER, O. H., & MOWRER ,W. M. Enuresis—a method for its study and treatment. *Amer. J. Orthopsychiat.*, 1938, **8**, 436–459.

MURRAY, E. J. The content-analysis method of studying psychotherapy. *Psychol. Monogr.*, 1956, **70** (13, Whole No. 420).

MUSSEN, P., & DISTLER, L. M. Masculinity, identification, and father-son relationships. *J. abnorm. soc. Psychol.*, 1959, **59**, 350–356.

MUSSEN, P., & DISTLER, L. M. Child-rearing antecedents of masculine identification in kindergarten boys. *Child Develpm.*, 1960, **31**, 89–100.

PAYNE, D. E., & MUSSEN, P. H. Parent-child relationships and father identification among adolescent boys. *J. abnorm. soc. Psychol.*, 1956, **52**, 358–362.

PETERS, H. N. Multiple choice learning in the chronic schizophrenic. *J. clin. Psychol.*, 1953, **9**, 328–333.

PETERS, H. N., & JENKINS, R. L. Improvement of chronic schizophrenic patients with guided problem-solving motivated by hunger. *Psychiat. Quart. Suppl.*, 1954, **28**, 84–101.

RACHMAN, S. The treatment of anxiety and phobic reactions by systematic desensitization psychotherapy. *J. abnorm. soc. Psychol.*, 1959, **58**, 259–263.

RAYMOND, M. S. Case of fetishism treated by aversion therapy. *Brit. med. J.*, 1956, **2**, 854–857.

RAZRAN, G. Stimulus generalization of conditioned responses. *Psychol. Bull.*, 1949, **46**, 337–365.

ROBINSON, NANCY M. Paired-associate learning by schizophrenic subjects under conditions of personal and impersonal reward and punishment. Unpublished doctoral dissertation, Stanford University, 1957.

Rogers, C. R. *Client-centered therapy.* Boston: Houghton Mifflin, 1951.

Rogers, C. R. Group discussion: Problems of controls. In E. H. Rubinstein & M. B. Parloff (Eds.), *Research in psychotherapy.* Washington, D. C.: American Psychological Association, 1959.

Rogers, J. M. Operant conditioning in a quasi-therapy setting. *J. abnorm. soc. Psychol.*, 1960, **60**, 247–252.

Rosenthal, D. Changes in some moral values following psychotherapy. *J. consult. Psychol.*, 1955, **19**, 431–436.

Rotter, J. B. *Social learning and clinical psychology.* Englewood Cliffs, N. J.: Prentice-Hall, 1954.

Rubenstein, E. A., & Lorr, M. A comparison of terminators and remainers in outpatient psychotherapy. *J. clin. Psychol.*, 1956, **12**, 345–349.

Rutherford, B. R. The use of negative practice in speech therapy with children handicapped by cerebral palsy, athetoid type. *J. speech Dis.*, 1940, **5**, 259–264.

Salzinger, K. Experimental manipulation of verbal behavior: A review. *J. gen. Psychol.*, 1959, **61**, 65–94.

Sarason, Barbara R. The effects of verbally conditioned response classes on post-conditioning tasks. *Dissertation Abstr.*, 1957, **12**, 679.

Schmidberg, Melitta. Psychotherapy of juvenile delinquents. *Int. ment. hlth. res. Newsltr.*, 1959, **1**, 1–2.

Sears, Pauline S. Child-rearing factors related to playing of sex-typed roles. *Amer. Psychologist*, 1953, **8**, 431. (Abstract)

Sears, R. R. Identification as a form of behavioral development. In D. B. Harris (Ed.), *The concept of development: An issue in the study of human behavior.* Minneapolis: Univer. Minnesota Press, 1957.

Sheehan, J. G. The modification of stuttering through non-reinforcement. *J. abnorm. soc. Psychol.*, 1951, **46**, 51–63.

Sheehan, J. G., & Voas, R. B. Stuttering as conflict: I. Comparison of therapy techniques involving approach and avoidance. *J. speech Dis.*, 1957, **22**, 714–723.

Shoben, E. J. Psychotherapy as a problem in learning theory. *Psychol. Bull.*, 1949, **46**, 366–392.

Skinner, B. F. *Science and human behavior.* New York: Macmillan, 1953.

Skinner, B. F. Some issues concerning the control of human behavior. *Science*, 1956, **124**, 1057–1066. (a)

Skinner, B. F. What is psychotic behavior? In, *Theory and treatment of psychosis: Some newer aspects.* St. Louis: Washington Univer. Stud., 1956. (b)

Solomon, R. L., Kamin, L. J., & Wynne, L. C. Traumatic avoidance learning: The outcomes of several extinction procedures with dogs. *J. abnorm. soc. Psychol.*, 1953, **48**, 291–302.

Sullivan, H. S. *The interpersonal theory of psychiatry.* New York: Norton, 1953.

Thirmann, J. Conditioned-reflex treatment of alcoholism. *New Engl. J. Med.*, 1949, **241**, 368–370, 406–410.

Thompson, G. N., & Bielinski, B. Improvement in psychosis following conditioned reflex treatment in alcoholsism. *J. nerv. ment. Dis.*, 1953, **117**, 537–543.

Tilton, J. R. The use of instrumental motor and verbal learning techniques in the treatment of chronic schizophrenics. Unpublished doctoral dissertation, Michigan State University, 1956.

Ullman, L. P., Krasner, L., & Collins, Beverly J. Modification of behavior in group therapy associated with verbal conditioning. *J. abnorm. soc. Psychol.*, in press.

Voegtlen, W. L. The treatment of alcoholism by establishing a conditioned reflex. *Amer. J. med. Sci.*, 1940, **119**, 802–810.

Wallace, J. A. The treatment of alcoholics by the conditioned reflex method. *J. Tenn. Med. Ass.*, 1949, **42**, 125–128.

Weide, T. N. Conditioning and generalization of the use of affect-relevant words. Unpublished doctoral dissertation, Stanford University, 1959.

Whiting, J. W. M. The research program of the Laboratory of Human Development: The development of self-control. Cambridge: Harvard University, 1954. (Mimeo)

Williams, C. D. The elimination of tantrum behaviors by extinction procedures. *J. abnorm. soc. Psychol.*, 1959, **59**, 269.

Williams, R. I. Verbal conditioning in psychotherapy. *Amer. Psychologist*, 1959, **14**, 388. (Abstract)

Wolpe, J. Reciprocal inhibition as the main basis of psychotherapeutic effects. *AMA Arch. Neurol. Psychiat.*, 1954, **72**, 205–226.

Wolpe, J. *Psychotherapy by reciprocal inhibition.* Stanford: Stanford Univer. Press, 1958.

Wolpe, J. Psychotherapy based on the principle of reciprocal inhibition. In A. Burton (Ed.), *Case studies in counseling and psychotherapy.* Englewood Cliffs, N. J.: Prentice-Hall, 1959.

Yates, A. J. The application of learning theory to the treatment of tics. *J. abnorm. soc. Psychol.*, 1958, **56**, 175–182.

(Received January 21, 1960)

Psychological Bulletin
1961, Vol. 58, No. 2, 160–174

THEORY OF SITUATIONAL, INSTRUMENT, SECOND ORDER, AND REFRACTION FACTORS IN PERSONALITY STRUCTURE RESEARCH

RAYMOND B. CATTELL
University of Illinois

Exploration of personality by multivariate experimental methods, as a means of objectively determining personality structure, has revealed, on the one hand, an array of stable, meaningful, cross-checking structures (Cattell, 1946, 1957; French, 1953), and on the other, some baffling inconsistencies. The latter have recently been pointed out by Becker (1960), apparently in criticism of the present writer's personality theory, but have been known for several years, and were, in fact, first brought to light by Cattell and Saunders (1950). Nevertheless, Becker does a service to advertise these facts; for psychologists have greatly neglected the solution of the problems revealed in this field.

The present writer's theoretical position is that it is conceptually correct to speak of the same unique source trait, e.g., cyclothymia-schizothymia, anxiety, ego-strength, surgency-desurgency, as something expressing itself (in terms of recognizable, replicable factor patterns) across all three possible media of experimental observation. That is to say, the same influence should appear in L data (life record, behavior *in situ*), Q data (questionnaire, consulting room, verbal self-evaluation), and T data (objective, laboratory, miniature situational, non-self-evaluative test performances).

In the article (Becker, 1960) to which I reply the fact that the actual correlation between the L-data and Q-data estimates of what are apparently equivalents in the two media, sometimes falls far short of perfection, is accepted as disproof of this theory. This theoretical conclusion is unsubtle; and the thesis of my reply is that countless threads of evidence contribute to the view that the same abstract personality source trait commonly operates across different media. However, certain "perturbations" have to be recognized which prevent the simple relation appearing on the surface, and these need to be taken into account in understanding psychological measurement generally.

In this area of scientific investigation, Becker has not asked the right question. Unexpected, but systematically evaluated perturbations of existing laws have often led to new discoveries, not so much by rejecting a law as by extending it, e.g., in astronomy in the discovery of Neptune through observed perturbations in the expected orbit of Uranus. So here, it is argued that there is no reason to abandon the notion of unitary source traits (Cattell, 1946) but that one must recognize certain new concepts, which we have introduced under the terms situational, instrument, and refraction factors. These are supported partly by marshaling existing evidence, but also by experiments undertaken ad hoc, but which, through an editorial veto on space to reply, have been reported in a separate publication (Cattell, 1960).

The Definition of Instrument Factors

The first and major source of perturbation in transmedia factor matching arises from what may be called *instrument factors*. Apparently, the first explicit recognition and demonstration of an instrument factor occurred in a structural analysis of a very widely selected set of objective personality tests, by Cattell and Gruen (1955), where a factor appeared literally produced by diurnal variations of sensitivity of a brass instrument (GSR). This purely instrumental influence created a factor by throwing common variance into all types of personality measures in which it was used. Such factors have appeared since in publications by Holzmann and Bitterman (1956), F. L. Damarin, D. T. Campbell, and L. Berwyn (unpublished), and several other unpublished studies known to the writer. Indeed, wherever questionnaire variables are mixed with ratings, attitude scales with questionnaires, or, sometimes, even one type of answer form with another, one or more factors may generally be found covering *all variables having formal similarity*.

The difficulty factors of Wherry and Gaylord (1944), and Dingman (1958), should definitely be regarded as a subspecies of instrument factor. Recently, in a study of the Music Preference Test of Personality (Cattell & Anderson, 1953) by Mayeske (1961) an instrument factor appeared even separating all items resting on one form of musical recording from those based on another technique. Instrument factors have become better understood in the last couple of years through extensive studies of their appearance in objective motivation structure analyses (Cattell, Radcliffe, & Sweney, 1960; R. B. Cattell & J. Horn, unpublished). There they appear as "vehicle factors" covering all objective devices using the same vehicle, e.g., information, autism, for the objective measurement of motivation strength. In this, and many similar contexts, it has been shown that instrument factors can be fairly clearly eliminated by ipsative scoring (R. B. Cattell & J. Horn, unpublished, see Table 1).

Before proceeding beyond this introduction by illustrations, to a more comprehensive definition of the concept of instrument factor, it is desirable, however, to make clear which peripheral factors are *not* to be included. This can be done most compactly by Figure 1, presenting a hierarchy which will be clear to multivariate experimentalists. Incidentally, the term "artifactors" is due to Roberts (1959), and has been sharpened by additional conditions here to make their separation from instrument factors cleaner.

The justification for the labels of the three forms of "perturbing" factors reproducible across experiments (matrices) will be given as we proceed. Concentrating first on instrument factors, let us note that they are definable, initially, only in terms of intention and perspective. Later, the definition can be made more satisfactory as we develop precise concepts indicating various universes of variables. For a quality which persists across the differences of content of a series of opinionnaires of similar form, and which perhaps consists of response to a particular form inherent in this instrument, though irrelevant to the content interest of the experimenter may yet represent behavior dependent on a real per-

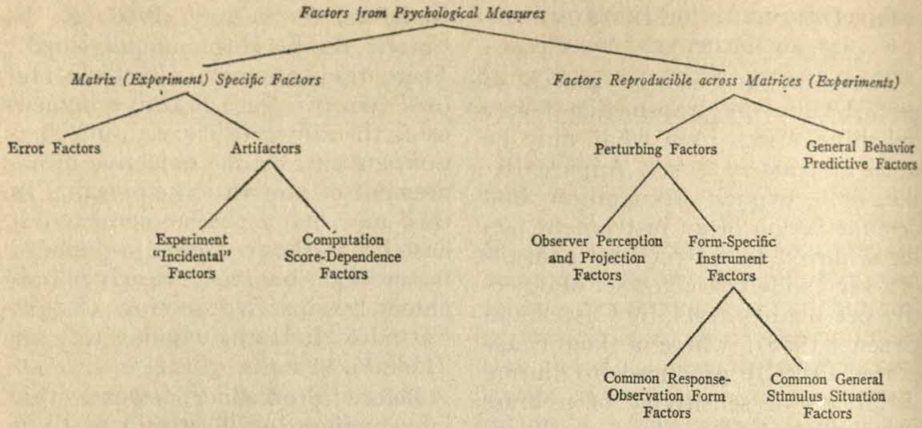

Factors from Psychological Measures
Matrix (Experiment) Specific Factors
Factors Reproducible across Matrices (Experiments)
Error Factors
Artifactors
Perturbing Factors
General Behavior Predictive Factors
Experiment "Incidental" Factors
Computation Score-Dependence Factors
Observer Perception and Projection Factors
Form-Specific Instrument Factors
Common Response-Observation Form Factors
Common General Stimulus Situation Factors

FIG. 1. The place of instrument factors in a taxonomy of factors.

sonality trait. For example, what comes as an instrument factor covering the variables of similar form, $a_1 \ldots a_n$, may well load (when $a_1 \ldots a_n$ are condensed to a single variable a, set in the new context of variables b, c, d, etc.) some important general personality factor.[1]

There is thus a sense in which an instrument factor is a matter of perspective, i.e., of one's starting point and of the plane of experience from which one chooses the majority of one's tests. In this sense, just as dirt is only "matter in the wrong place," so an instrument factor is only "variance where we didn't expect it or don't want it." When we are measuring personality by questionnaire we obviously do not want *each and all* of the diverse personality dimensions included to be contaminated by what might be called a "generalized specific," i.e., a specific to questionnaires. And the fact that that specific may, indeed, be something more than a trivial specific, but an expression of a single important personality factor spread over and contaminating all the alleged diverse personality measures, does not make the measurement harsh any more acceptable!

When more progress has been made toward a systematic taxonomy of tests, on some such objective basis as that worked out by Cattell and Warburton (in press), it would become possible to set up also a relatively objective classification of instrument factors, according to the types of personality approach to which they are tangential. For "form" and "content" are quite

[1] Incidentally, it is the failure to recognize this perspective which, in the present writer's opinion, has made so much recent work on response sets a rather uneconomical use of psychological research time. Whereas educational psychometrists during the late 1950s "discovered," in their opionnaire tests, response sets (Cronbach, 1950), social desirability sets (Edwards, 1957), extremity of response sets (Berg, 1955), and acquiscence—tendency to agree, yes-vs.-no (Messick & Jackson, 1961)—these had already been employed by designers of objective personality tests in the late 1940s and early 1950s (Cattell, 1946; Cattell & Gruen, 1955). In the context of broader personality theories, and varied behavioral measures involved, it had already become clear that what itemetrists, without knowledge of the literature in this area, later treated merely as "flaws" in their paper-and-pencil tests, were actually expressions of well defined personality factors, e. g., anxiety or UI 24, comention or UI 20, superego rigidity or UI 29, as well as UI 31 (Cattell, 1957).

subjective categories, and, in any case, by no means exhaust the possible planes of experiment to which instrument factors can be orthogonally intrusive. For the time being, however, we must take a relativistic position, and one centered in "content." On this basis we shall contingently define an instrument factor as any uniquely (simple structure) rotated factor which covers a whole set of diverse variables having formal resemblance in presentation, mode of permitted response, or scoring, and which does not extend to tests of the same psychological content when couched in other modes of formal presentation, response, etc.

Theory of Sources of Perturbation Affecting Trait Alignment

It should be noted that there are two distinct, though related senses in which a source trait can be said to be the same or not the same in two different media:

1. An estimate of the factor from the variables in one medium may correlate less than unity with its estimate from variables in another medium, even when attenuation-corrected for (*a*) unreliability of measurement, and (*b*) imperfection of estimate.

2. It may not be possible to discover a trait, when factoring both media together, which has simple structure across both media and also possession of the hypothesized, similar-meaning salient loadings in both media. (Whether one also means that the simple structure position in one medium will not project into the other we shall discuss below.)

Becker has been concerned with the first of these, denying alignment without first checking that Corrections *a* and *b* could not restore the correlation to unity. In any case, the second meaning is more important. If unity in this sense holds, personality theory is profoundly simplified, and it is only a matter of the mechanics of statistics to produce weighted measures from the two media that *will* approach a correlation of unity.[2]

In the larger collation of data and new experimental work (Cattell, 1960), from which the present article abstracts, it has been shown that the presence of unrecognized instrument factors in the two media will prevent alignment either in Sense 1 or 2, unless special new techniques are used. Before devoting a section to closer inspection of this result, however, it is desirable to set out a clear theory about more general sources of perturbation. For, in principle, one can see that there are some six possible origins of the failure to find a one-to-one alignment of primary personality factors measured in one medium with those measured in another. Some of these will produce instrument factors; others will contribute to other kinds of nonalignment to be described.

Sources of Nonalignment

Human transmission (perception, evaluation, projection, memory) of score values. Largely this means rating and self-rating (L and Q data). This is too subtle and complex a field—hitherto handled too simply in terms

[2] Such a procedure should be sharply distinguished from what Becker (1960) appears to advocate, and describes Gough as doing, namely, to force a Q scale to align itself with an L factor by assiduous item selection. Any such procedure contributes nothing to our knowledge of structure, but only hides the problem. If it succeeds, and if our theory is correct that L-data factors are the most heavily contaminated of any medium with irrelevant factors, this is forcing a poorly oriented measure to agree with a still poorer one.

of "halo"—for the present abstract summary to be illustrated in available space (see Cattell, 1960). Theoretically, the pattern of correlations, and therefore of obtained factors, could be distorted by, and only by, properties of the individual and his relation to the recorder which affect the recording of *all* his behavior variables, and by properties of the perceiving recorder. The former can be divided into (*a*) value relationships, of which liking-disliking (a constituent in halo) is only one; and (*b*) perspicacity or visibility effects, e.g., extraversion making the ratee more known, position effects making certain behaviors more clear. The latter can be divided into projections of (*a*) stereotypes or cultural clichés,[3] and (*b*) refraction factors, discussed below, peculiar to one medium. In all the "perceiving recorder effects" a correlation is produced by "projection" of a (perhaps quite unconscious) conviction that certain variables go together. Some of these may produce typical instrument factors, uniformly and about equally loading all variables in the medium; but others may load only *some* variables, producing what are perhaps best described as "perception-evaluation" projection factors, and which are not true instrument factors.

Communality of variables in respect to some trait required for handling a similar formal performance in all of them (*or for registering in an observation situation*). This is essentially one of the two main sources (see following paragraph) of instrument factors only. The countless possibilities may be illustrated by e.g. the use of 30 scales in all of which the score (in one direction or the other) depends on an ability to read, or on information or skill of expression, or tendency to say yes rather than no, etc.

Communality of variables in respect to scoring or scaling applied after administration. Quite apart from common demands on the subject's actual performance as in the previous paragraph, anything in the formal *scoring* procedure which tends to give similar sigmas, and skewedness (and in some cases means) throughout one class of tests will tend to create higher correlation among them and a common factor. That is to say, if the matrix of correlations of tests a_1 through a_n were just the same, on a rank formula, as that for b_1 through b_n, but if all the a's, on the one hand, and all b's, on the other, have similar distribution, then basing the matrix afresh on a product-moment formula will tend to give an instrument factor for the a's and/or the b's separately.

Coincidence of different global stimulus situations with different test media administrations. If a person answered one set of questionnaires in private, and another orally and publicly (which is akin to the interview or behavior rating situation), we should expect real differences in response due to the actual stimulus situation, covering the occasion on which all items of one test were answered, being different from that covering the other test-taking setting. A priori this could create both an instrument factor, conterminous with each medium-situation, and also a change in loading of the same items on the same personality factors in the two situations.

Habitual broad area differences in actual trait development and expression. Among children, for example,

[3] Since sociologists have ruined "stereotype," by applying it equally to a widespread concept which either (*a*) does or (*b*) does not, correspond to statistical reality, I suggest "cultural cliché" explicitly for a widespread cultural concept which is significantly different from any externally existing pattern.

we should expect the particular behavior variables representing, say, the dominance factor, to be expressed to different degrees in the home environment and in the school environment. This is analogous to the point in the above paragraph, except that the influence is expressly conceived not to lie in the temporary measurement situation itself, but in the prolonged life situation, leading to real differences of actual habit strength, i.e., of the trait itself. Factor analytically, this might produce a home dominance factor and a school dominance factor, representing the relative impact of home and school, respectively, or alternatively, one factor modified by two other factors, each peculiar to one broad area. If the former proves to be more characteristic, then we can confidently predict that the two first-order factors will correlate highly and yield a single second-order dominance factor. Even if the former is true it would be possible, in a rough factoring to perceive the structure as that of a home and school instrument factor (as in the second possibility) but psychologically, the interpretation, if the proper structure is obtained, would now be different from an instrument factor effect. The area differences would then be interpreted as *real* structure differences, and the concept of a single dominance trait would be discovered and justified *only* at the second-order factor level.

Differences among media in density of representation of variables. If in sampling variables in the ability field an experimenter accidentally took one variable for each of Thurstone's primary abilities and factored, he would obtain, straightaway, i.e., as a first-order factor, that general ability factor which, in any "dense" representation of variables, appears only as a second-order factor (Thurstone, 1938). This concept of density of variable representation has been developed further elsewhere (Cattell, 1957, pp. 808–817), but it is easy to see that if there were really large differences of density unrecognized between media we should obtain no correlational alignment of the primaries in the two fields. Only on exploring the second order would the possibility arise of discovering that a second order in one medium is the same as a first order in the other.

Actually, as soon as systematic exploration of second-order structure in questionnaires reached to six factors (Cattell, 1957; Cattell & Scheier, 1961; Cattell & Warburton, in press), it became evident that four second-order *questionnaire* factors aligned with four first-order *objective test* factors (UI 19, 20, 24, and 32); and in two of these, UI 24 (anxiety) and UI 32 (extraversion), the agreement is perfect within small limits of experimental error. An instance from a different realm, but amounting to a correlation of only 0.80 between the two media, exists in Tollefson's demonstration (1961) that the second-order extraversion factor in the questionnaire is a first-order factor in the Humor Test of Personality. These alignments (from the earlier, 1954–1957, publications above) are not mentioned in Becker's article (1960), perhaps because his comments are all on L- and Q- (rather than T-) data alignments. But the findings are highly relevant as showing that there does exist a corner of the intermedia jigsaw puzzle which is beginning to fit in place. These five experimental instances alone are surely sufficient to encourage us in that rejection of nihilism which this article undertakes.

To risk a prediction in the little explored field of "density," one might judge that variables in Q data will prove somewhat more "dense" than L data. But substantially, as the above evidence shows, one can conclude only that variables as commonly chosen are much more dense in Q than T data. This is understandable; e.g., in the T-data anxiety factor, we test startle response by a single cold pressor test (Cattell & Scheier, 1961) whereas in most anxiety questionnaires there are a dozen items asking in different ways how easily the person startles. Cronbach (1960), Comrey, and others who criticize low homogeneity when reviewing factor scales, are perhaps unwittingly driving their flocks toward the more serious danger of using personality scales heavily loaded in spurious "specific" variance of this latter kind, instead of watching that their scales deal with personality factors having broad psychological relevance and effectiveness.

If the above search for sources of perturbation has been truly exhaustive, our summary must include three other forms of distortion besides instrument factors, constituting four in all, as follows (beginning with instrument factors):

1. Test instrument factors, including common test form (response-observation-score) factors, and common test general stimulus situation factors.

2. Modification of actual trait by influences peculiar to one area of expression, producing primaries for each area and requiring conceptual unity to be sought at a higher order level.

3. Difference of density of representation of variables, as commonly unconsciously chosen by experimenters, in their different media, resulting in a higher order in one medium matching a lower order in another.

4. Perception-evaluation or projection factors, which trespass on the variance of the variables used to estimate personality factors, *not* by uniformly loading all in one medium (as does an instrument factor) but having each a characteristic form, and, when restricted to one medium, having the properties of refraction factors described below.

The Practical Problem of Reaching Personality Structure Despite Distortions

If the above theoretical analysis is correct the manifest correlational picture of personality structures will be less like Whistler's portrait of his mother than the cubist's rendering of the same, fractured into surprising new supernumerary planes and facets. To translate from the latter to the former, it is necessary that research, first, check the hypotheses about the forms of distortion at work and, second, find experimental and statistical means for isolating and setting aside these various perturbing influences.

One cannot do more than glance at these tasks here. As to the first, our initial examination of data shows definitely that form-specific instrument factors exist, while my colleagues and I have also begun to give evidence for the Sources 2, 3, and 4. The source of nonalignment labeled 2—local area modification of real traits—has been more fully illustrated elsewhere (Cattell, 1960) but must be left to others systematically to investigate. Source 3, changing density with changing medium, has already been substantiated.

As to the second task—segregating the distorting influences to arrive at essential structure—the unraveling of Effects 2 and 3 above is straightforward, by second-order factoring, though the possibility has been mooted above that Source 2 could produce two instrument factors, beyond a single first-order factor, instead of two first orders resolving into a second.

Setting 2 and 3 aside, therefore, we shall devote the present section to unraveling the effect of instrument factors, 1 above, and the following section to perception-evaluation-project phenomena, 4 above.

The special experiments with instrument factors described elsewhere (Cattell, 1960) proceeded first to find what happens when one factors correlation matrices derived from known, numerically stated factor models, and secondly, to experiment with varieties of solution in actual psychological data where the existence and boundaries of an instrument factor were well known beforehand. These experiments showed that:

1. Where the instrument factor covers *all* variables, i.e., where they are not embedded in a larger matrix, with other media to constitute a hyperplane and determine unique rotation, the typical investigator and procedure will not find or be aware of the instrument factor.

2. If the instrument factor is not found then either: (*a*) the correlations among the primaries will be distorted (if it is positive on all and they are all positively correlated, it will increase their correlations); or, (*b*) the simple structure which really exists among the primaries will not be found, or found only in very impaired form. Commonly *b* will predominate, but both will operate.

After this demonstration of the effect of an instrument factor in a single medium we proceeded to models and real instances containing blocks of variables uniformly from each of two or three media. Herein each medium was covered by *one* instrument factor but where *true* personality factors existed in the sense of having a simple structure position with salient loadings on variables of similar meanings in *both* media. Here it was shown:

1. If one obtains the best possible simple structure (perhaps imperfect because of mixed-in instrument factor) among variables separately in each medium, the same simple structures cannot usually be found when the media are put together.

2. One reason for this is that if one projects the simple structure position satisfactorily obtained in one medium into the second,[4] it definitely does not give simple structure within the second.

3. If, however, one first admits the existence of, and locates by simple structure in the combined matrix, the instrument factors (which can now have determinate hyperplanes), then the true personality factors, operating across both media, can be located (in blind simple structure rotation). A successful example of this in real data—objective motivation measurement (R. B. Cattell & J. Horn, unpublished)—is shown in Table 1 here, and in other models elsewhere (Cattell, 1960). Our ignorance of this principle in 1948 was presumably responsible for the chaotic outcome of the first extensive transmedium factor analyses (Cattell & Saunders, 1950, 1955).

Incidentally, it will be obvious that missing the instrument factor, failing to rotate it correctly if one does not miss it, and encountering the subsequent distortion are due respectively to (*a*) the lack of a test for factor extraction that will decide,

[4] This cannot be done, of course, simply by applying the same discovered transformation (λ) matrix to the centroids, because the latter begin at different positions. One first discovers by the Procrustes program the λ most nearly reproducing the first medium simple structure from the joint medium centroid.

TABLE 1

PSYCHOLOGICAL AND INSTRUMENT FACTORS AS FOUND IN OBJECTIVE, DYNAMIC TRAIT SIMPLE STRUCTURE

Attitude Variable and Device Measurement	Factor Matrix				
	Psychological Factors			Instrument Factors	
	Escape Erg	Sentiment to Parents	Self-Sentiment	Information Device Factor	Autism Device Factor
1 Desire for good self-control. Information measure	00	−02	$\boxed{26}$	$\boxed{54}$	03
2 Wish to know oneself. Information measure	03	−05	$\boxed{31}$	$\boxed{27}$	19
3 Wish to never to become insane. Information measure	−06	12	$\boxed{22}$	$\boxed{43}$	04
4 Readiness to turn to parents for help. Information measure	−02	$\boxed{35}$	09	$\boxed{28}$	−01
5 Feeling proud of one's parents. Information measure	−06	$\boxed{28}$	−01	$\boxed{24}$	01
6[a] Desire to avoid fatal disease and accidents. Information measure	$\boxed{16}$	04	13	$\boxed{65}$	−02
7[a] Wish to get protection from A bomb. Information measure	$\boxed{14}$	−08	03	$\boxed{14}$	−05
8 Desire for good self-control. Autism measure	01	−04	$\boxed{30}$	02	$\boxed{22}$
9 Wish to know oneself. Autism measure	−08	07	$\boxed{37}$	−01	$\boxed{31}$
10 Wish never to become insane. Autism measure	00	−01	$\boxed{16}$	00	$\boxed{25}$
11 Readiness to turn to parents for help. Autism measure	−08	$\boxed{18}$	09	−08	$\boxed{42}$
12 Feeling proud of one's parents. Autism measure	−03	$\boxed{14}$	01	06	$\boxed{14}$
13[a] Desire to avoid fatal disease and accidents. Autism measure	$\boxed{26}$	20	01	04	$\boxed{17}$
14[a] Wish to get protection from A bomb. Autism measure	$\boxed{23}$	13	09	15	$\boxed{10}$

Note.—The theoretically required salients to define the factors are boxed in, and except for two values at the bottom of the parental sentiment factor column, the salients are high (above .09) where, and only where, they are theoretically required to be.

[a] Attitudes 13 and 14 are the same as 6 and 7, but in a different medium, and similarly, for the other cross-media personality factors.

to within less than an error of two or three factors, how many should be extracted; (*b*) having no variables from other media to give a hyperplane for it; and (*c*) the variance that should have been in the instrument factor being pushed into the personality factors, destroying the clarity of their hyperplanes. The remedy which worked in the above cases was to give good technical attention to these issues.

ON ISOLATING TRANSMEDIUM PERSONALITY FACTORS AND REFRACTION FACTORS

Our final step consisted in returning to the actual L and Q data from which Becker infers that personality factors are unmatchable across media, and showing that when examined by more penetrating concepts, as above, uniquely determinate, psychologically meaningful, factor patterns appear, expressing themselves appropriately in both media for each factor. This has theoretical interest in giving additional substance to Point 3 above, by introducing the notion of refraction factors, and in producing some order in that L-Q frontier which has hitherto been the most hopelessly obscure of the transmedia relationships. Nevertheless, this approach does no more than reveal *some* order, and at the same time opens the door on a lot of problems, particularly in the field of behavior rating, which will now demand systematic investigations.

It is not easy to find in any published study of the past 20 years (ever since personality structure research began in earnest) an experiment really adequate in reaching the technical conditions necessary to get anywhere on this question. One needs, among other things, an experiment: (*a*) on a sufficient sample for sampling errors not to be intrusive; (*b*) where the subjects had a long testing period in which they were simultaneously rated *in situ* and subjected to questionnaires, comprehensive, reliable, and valid enough to define several factors clearly; (*c*) where ratings and questionnaire variables were strategically chosen to represent psychologically *familiar* factors, already vouched for by earlier researches; and (*d*) where ratings were carried out by peers and under the requisite conditions described elsewhere (Cattell, 1946, 1957). Probably the most satisfactory data available is that in which the experimental work was broadly conceived and painstakingly carried out by Coan, on 7.8-year-old children (Cattell & Coan, 1957, 1958). It suffers only with respect to *d*, in that ratings were made by teachers instead of peers, and perhaps in reduced homogeneity of sample through equal inclusion of boys and girls.

Taking the data of this experiment we find that 24 rating variables have already been factored and blindly rotated into 12 very definite simple structure factors, each represented by two markers (see Table 5 in Cattell, 1960). Similarly, 24 variables in Q data, each consisting of a scale of about eight items, have been resolved as 12 well known simple structure factors, each marked essentially by two salient variables. However, on psychological inspection of these resolutions, the hypothetical position was taken that only 10 of the 12 factors were common to the two matrices, the remaining 4 being special, 2 to each matrix.

The two sets of 24 variables were now combined and intercorrelated in a cross-medium, L-Q matrix of 48 variables, which, by Tucker's test, yielded 16 factors. (With the hypothesis of matching, above, one would expect 14, but it is usual to find some new factor created by the mixture when two matrices are pooled.) The structure of this new factor space proved to be complex. Projection of simple structure obtained in one into the other, as described earlier (Footnote 4), would not yield a good combined simple structure. Attempts to force simple structure by varimax, oblimax, or other "analytical" programs failed because these rigid programs could not recognize and uniquely rotate the instrument factors, which, on the basis of the above principles and findings, we knew must be present. Only a patient and comprehensive exploratory visual rotation (aided by the photographic Rotoplot program on Illiac), over 22 rotations, yielded a position of such stability that one could repeatedly return to it. In reaching this position we found that the hyperplanes in the data were noticeably a little broader (about ±.13 instead of ±.10) than those existing in one medium alone.

On examining the solution, set out in Table 2,[5] we found that we had essentially an instrument factor for L data and another for Q data (not set out at the *end* of the matrix, but marked In_L and In_Q, in Table 2). There are also two other factors, which we would guess might be projected "clichés," numbered 13 and 16. The interesting fact is that when this debris is set aside, patterns for the well known personality dimensions C (Ego strength), D (Excitability), F (Surgency), and H (Parmia), appear, with the appropriate four markers (2L and 2Q) on each, though the hyperplanes are pierced by one or two random appreciable loadings on other factors. (Counting within ±.13 they reach acceptable percent-

[5] The matrix containing the correlations among factors, the lambda matrix, and the centroid for Table 2 have been deposited with the American Documentation Institute. Order Document No. 6570 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance $1.75 for microfilm or $2.50 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

TABLE 2

Simple Structure Rotation of Combined L and Q Data, with Regard for Instrument Factors

(Primary Factor Pattern)

		C	D	F	H	In$_L$	In$_Q$	A$_Q$	G$_L$	G$_Q$	J$_Q$	C$_L$	O$_{40}$	13	O$_Q$	D$_Q$	16
L	A	+20	−03	+05	+06	+44	+09	−10	−07	−16	+04	−20	+06	−48	+02	−01	−25
	C	$\boxed{+45}$	−13	−15	−03	+04	−01	−19	−04	+05	+04	−64	−11	−16	−08	+00	−03
	D	+07	$\boxed{+45}$	+00	−12	+67	−03	−11	−10	−03	+15	+02	+04	+07	−03	−10	−10
	E	+13	−31	−08	+30	+43	+03	+09	−10	+01	−09	−16	+21	+04	+04	−04	+22
	F	+39	−05	$\boxed{+50}$	−09	−22	−01	+10	+06	+07	+04	+14	+08	+07	−01	−07	−25
	G	−19	−09	+01	+08	−19	+01	−05	$\boxed{+66}$	+00	−01	−01	−11	+05	−04	−07	+00
	H	+10	−03	−07	$\boxed{+48}$	−11	−06	−01	+50	+08	−01	−01	−04	+06	−22	+04	−07
	I	−14	+10	−01	+05	−56	−21	−10	−13	−11	+00	+06	+40	−07	+00	−10	+19
	J	−04	+31	−20	+09	+45	+09	−35	−04	+00	−08	+00	−33	−04	−09	−10	−05
	L	+00	+19	+00	+14	+60	−02	+00	−03	+09	+05	−12	−65	+02	−04	−09	+08
	M	+10	−08	−07	−17	−22	−03	+16	+70	+01	−01	+06	+19	−02	+06	+13	−09
	O	−19	−09	+07	−44	−11	−18	+11	−38	+12	−09	−02	−10	−53	+03	−12	+15
	A	+03	−28	+15	+53	+18	+15	+06	+00	+11	+04	−02	−08	−10	+03	−19	−02
	C	$\boxed{+49}$	−12	+19	+19	−22	+03	+09	+07	−01	+08	−56	+08	+11	+10	+04	−06
	D	+00	$\boxed{+20}$	+00	−14	+84	+00	−05	−10	+09	+04	+12	−06	+07	−07	−13	−13
	E	+00	+05	+01	+48	+71	+02	+07	+19	+03	+05	+10	+18	−10	+08	−19	−01
	F	−13	+03	$\boxed{+46}$	+27	−21	−10	+06	−65	−04	−09	+10	−01	+02	+02	−07	+08
	G	+12	+12	+07	+25	−19	+07	−01	$\boxed{+76}$	−04	−05	−08	−04	+13	+00	+06	+11
	H	−07	−09	+04	$\boxed{+55}$	−21	+02	+06	+22	+07	−02	−04	−08	+08	−08	−23	+01
	I	+32	−01	−07	+09	+05	+12	+05	−22	+03	−05	−07	+84	+13	+01	−03	+05
	J	−06	+12	−04	−41	+84	+07	+04	+07	+01	−05	−10	−03	−02	+02	−27	−10
	L	+10	+06	−11	+00	+67	−03	+07	+10	+00	−13	−35	−06	+02	−01	−24	−14
	M	+12	+01	−07	+44	+40	+03	−05	+37	+04	+04	+12	+04	−42	−02	+27	+01
	O	−03	−22	+07	−48	−34	+07	−14	+13	+07	−07	+20	−00	−16	−01	+13	+30
Q	A	−03	+06	−19	+06	+10	+01	$\boxed{+58}$	−04	+07	−13	−10	−51	+01	+03	+10	−09
	C	$\boxed{+36}$	+28	+07	+20	+05	−03	+12	+07	+03	−06	+24	−17	+04	+04	−12	+06
	D	+01	$\boxed{+65}$	+00	+20	−11	+08	+02	−04	+36	−05	−12	+06	−10	+03	−49	−05
	E	−27	+14	+00	−03	−01	−07	+02	−18	−51	+01	+04	−04	−33	−03	−26	+10
	F	+00	+10	$\boxed{+49}$	−03	+15	+45	+04	−06	+04	−07	−14	−17	−13	−11	+00	+29
	G	−16	−01	−05	+03	+11	−31	+00	−04	$\boxed{+58}$	+02	+06	+03	−04	−17	−04	−01
	H	−10	−06	−09	$\boxed{+36}$	−12	+03	+06	+10	+01	+06	+02	−04	−10	+37	+58	−01
	I	−17	−13	−07	+02	−01	+72	+00	+00	+08	+12	+13	−11	+00	−07	+00	−07
	J	−12	+10	−05	+05	−07	−07	+10	+01	+05	$\boxed{+66}$	+02	−06	+05	−34	+03	−10
	N	+19	−09	−27	−17	−22	−26	$\boxed{-47}$	−25	+05	−07	−13	+12	−15	−02	−14	+02
	O	−04	+04	+03	+09	+03	−64	−16	+04	−08	+05	−01	−00	+08	+30	−04	+06
	Q$_4$	−29	+06	−01	+09	−08	−12	+10	−03	+04	−05	+02	$\boxed{+40}$	+01	−12	+03	−33

	C	D	F	H	In_L	In_Q	A_Q	G_L	G_Q	J_Q	C_L	O_{40}	13	O_Q	D_Q	16
A	+14	+06	−35	−08	−01	+01	$\boxed{+51}$	+04	−08	−07	−16	−18	+00	−01	−12	+06
C	$\boxed{+55}$	+06	−01	−05	+04	+02	+05	−04	−05	+05	+08	−04	−11	+03	+00	+03
D	+12	$\boxed{+60}$	+04	−05	−26	−12	+10	+09	+44	+02	−18	+19	+11	+10	−42	+11
E	+04	+40	−04	−14	−01	−03	+05	−16	+00	−29	+10	−07	−23	+02	−58	−03
F	−07	+15	$\boxed{+39}$	−25	+10	+06	−02	+03	−07	−04	−07	+08	+04	+02	−12	+08
G	+09	+17	−03	−11	−05	+30	−06	−06	$\boxed{+80}$	+06	+04	−18	−11	−03	−12	−06
H	−06	+08	+11	$\boxed{+69}$	−11	−06	−02	−09	+01	−07	−04	−04	+01	−04	+52	+06
I	−06	−01	+01	+06	−11	+70	+19	+13	+49	−12	+05	−10	+05	+12	+06	+05
J	+12	+09	+01	+05	+07	+01	+00	−06	−03	$\boxed{+45}$	−05	+01	−10	+21	−12	+17
N	+13	+04	+07	+06	+05	−18	$\boxed{-42}$	−03	+16	+04	−10	−04	+12	+15	+06	−02
O	−17	+03	−08	+03	+12	−31	+01	−01	+13	−05	+13	−12	−01	+46	−17	−31
Q_4	−06	+00	+04	+03	+01	−06	+01	−06	+11	+01	−11	$\boxed{+57}$	−07	+02	+17	−37

ages of 65, 73, 77, and 54 in the hyperplane.)

However, a hitherto undescribed phenomenon is encountered here, namely, the appearance of factors restricted to one medium, *and appearing in one or both of the separate media alongside, and simultaneous with, the appearance of the joint medium factor having the same personality meaning.* This is illustrated by C and C_L, D and D_Q (Table 2), wherein the real psychological factor (C or D), loading the four essential variables across both media, carries alongside it an incomplete image of itself in each medium. The incomplete image loads only the two variables which belong in one medium. To these patterns, occurring simultaneously with the combined pattern, I have tentatively given the name "refraction factors," since they are analogous to what would be seen if one looked at an object both directly and refracted through a prism of another medium, one on each side of the line of vision.

Actually Table 2 does not simultaneously present *all* refraction factors for all real factors, but this should not disturb us any more than the failure of a single archeological digging to provide all the bones of a skeleton or all cultural elements for a given period. For, as it has been argued elsewhere (Cattell, 1958) any matrix typically has strictly as many dimensions as variables, and probably even more hyperplanes, i.e., one is always taking a selection in simple structure among more possible hyperplanes than one has chosen to extract factors. Further search should be made for refraction factors, therefore.

A vital empirical question affecting further inference at this point concerns the correlations among the real and refraction factors for a given psychological dimension. We had expected them to be positively correlated, but the best estimate from existing data is that they are only slightly correlated, if at all. It is possible, however, that if more dimensions had been taken out their correlations would have been increased (see Diagram 5, Cattell, 1958).

Exploration and evaluation of possible hypotheses to account for refraction factors would require at least an article to itself. One does not go too far in interpretation, however, to say that they imply that each individual, in addition to his assessment on the real factor, gets a "bonus" on the variables peculiar to each medium, which is substantially unrelated to his status on the real factor. Our hypothesis is that these refraction factors belong to the perceptual class (Class 4 on page 166 above) and arise from the behavior in question being differently perceived in the two media. In self-rating a varying sensitivity and self-awareness—only in special cases a function of the trait being rated—could provide the differing "bonus" from person to person. The differing visibilities of these individuals from the position of the rater, giving the L-data refraction, would be expected to be quite unrelated to the order of their individual sensitivities in self-rating.

If this is correct one might also expect the lesser loadings, on variables other than the two salients, to be systematically different on the two refraction factors. For example, the rating by others, in the case of a factor much concerned in delinquency, might impart something of the stereotype of a scoundrel, where the Q-data refraction factor might convey more of a good person in difficulties. Since our main concern is with the order which emerges little has been said of the "debris" factors notably 13 and 16 in Table 2. But our conclusion, tentatively, is that "evaluative" and "visibility" factors other than refraction factors are at present run together in the insufficient factor space so far used, and that, especially in the L data, these "halo" and related factors are substantial. They do not appear to be any known second-order factors, which can sometimes appear in inadequate first-order factorings. It has sufficed for our present investigation simply to set them aside. But if closer research scrutiny in this heap shows that our present indications are correct that these Class 4 perturbers are much larger in L than Q data, then the practice of trying to force questionnaire factors to align with rating "criteria" comes still more in question than it is today.

That the reader may more directly evaluate the nature and quality of the simple structure in Table 2 we have set out in Figure 2 a plot of two psychological ("real") factors therefrom.

Summary and Conclusion

1. Correlations among primary personality factors in different media do not provide a simple pattern of one-to-one relations, and fall decidedly short of unity between two factors of the same apparent psychological meaning.

2. The theoretical possibilities and the natural occurrences of perturbing influences hiding true alignment have been discussed and demonstrated. They have been classified as (*a*) test instrument factors; (*b*) actual trait modification by differing experience in subareas, requiring unity to be sought at a higher order level; (*c*) differences of density of representation of variables in different media;

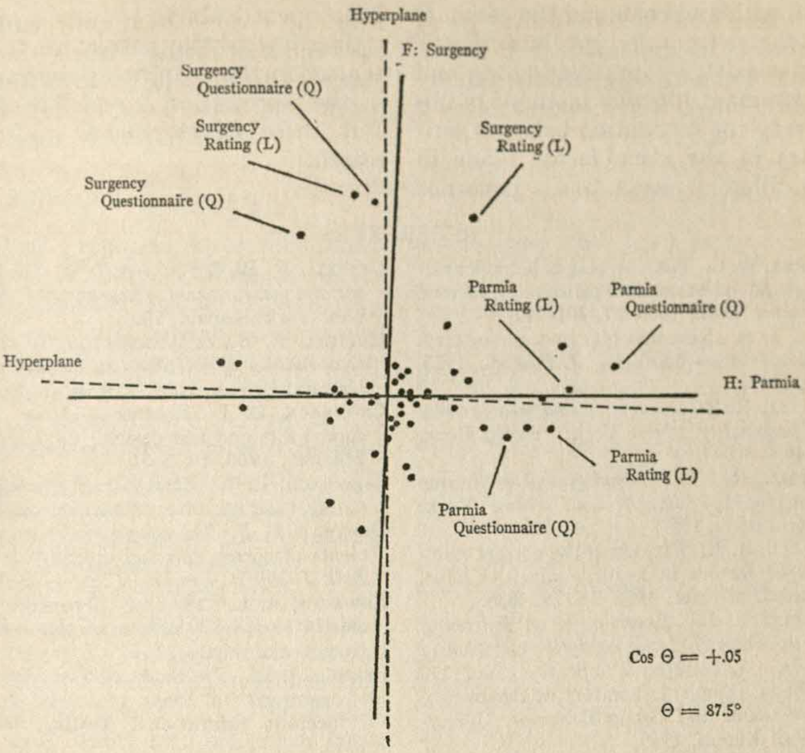


FIG. 2. Simple structure appearing between cross-media personality factors. (Marker variables labeled)

and (*d*) perceptual-evaluation-projection factors, occurring where human transmission of observations is involved.

3. Experimenters, especially when leaving rotation decisions to falsely founded analytical computer programs, commonly miss instrument factors, but when these are properly isolated and set aside by careful experiment it is possible to find the well known primary personality factors, each appearing as a single factor expressing itself in both L and Q media.

4. Regard for instrument and second-order–first-order factor relations is already producing clarity and consistency in personality structure research; but much remains to be explored regarding at least four forms of distortion which apparently occur where human transmission is involved, i.e., in L and Q data. The new phenomenon of refraction factors particularly calls for intensive research.

5. One must distinguish between the question "Does a single simple structure factor exist loading variables of the same meaning on both media?" and "Can one get a perfect correlation between estimates of apparently (by meaning) the same factor, made in the two media?" Even when the answer to the first, so important for personality theory, is "Yes," as this paper claims to have shown, the answer to the second remains "No." The variance due to instrument factors, refraction factors, and any evaluation-perceptual factors peculiar to one medium will re-

main with and confound the estimate of a factor from that medium. Possibilities exist, by ipsative scoring and discriminant function methods of improving the correlation between estimates of the same factor made in two different media, and a path has been opened above toward a proper estimation of the correction for attenuation that can be applied to see if the correlation could be unity. But these developments await research.

REFERENCES

BECKER, W. C. The matching of behavior rating and questionnaire personality factors. *Psychol. Bull.*, 1960, **57**, 201–212.

BERG, I. A. Response bias and personality: The deviation hpothesis. *J. Psychol.*, 1955, **40**, 61–72.

CATTELL, R. B. *Description and measurement of personality.* New York: World Book, 1946.

CATTELL, R. B. *Personality and motivation structure and measurement.* New York: World Book, 1957.

CATTELL, R. B. Extracting the correct number of factors in factor analysis. *Educ. psychol. Measmt.*, 1958, **18**, 791–838.

CATTELL, R. B. *Experiments on sources of perturbation in factor analytic resolution of traits.* (Advanced Publication No. 11) Urbana, Illinois: Laboratory of Personality Assessment and Group Behavior, University of Illinois, 1960.

CATTELL, R. B., & ANDERSON, J. C. *The IPAT Music Preference Test of Personality.* Champaign, Illinios: Institute for Personality and Ability Testing, 1953.

CATTELL, R. B., & COAN, R. W. Child personality structure as revealed in teachers' behavior ratings. *J. clin. Psychol.*, 1957, **13**, 315–327.

CATTELL, R. B., & COAN, R. W. Personality dimensions in the questionnaire responses of six and seven year olds. *Brit. J. educ. Psychol.*, 1958, **28**, 232–242.

CATTELL, R. B., & GRUEN, W. The primary personality factors in eleven year old children by objective tests. *J. Pers.*, 1955, **23**, 460–478.

CATTELL, R. B., RADCLIFFE, J., & SWENEY, A. The objective measurement of motivation structure in children. *J. clin. Psychol.*, 1960, **16**, 227–232.

CATTELL, R. B., & SAUNDERS, D. R. Interrelation and matching of personality factors from behavior rating, questionnaire and objective test data. *J. soc. Psychol.*, 1950, **31**, 243–260.

CATTELL, R. B., & SAUNDERS, D. R. Beiträge zur Faktoren-Analyse der Personlichkeit. *Z. exp. angew. Psychol.*, 1955, **7**, 319–343.

CATTELL, R. B., & SCHEIER, I. H. *The meaning and measurement of neurosis and anxiety.* New York: Ronald, 1961.

CATTELL, R. B., & WARBURTON, W. W. *A compendium of objective tests in personality.* Urbana: Univer. Illinois Press, in press.

CRONBACH, L. J. Further evidence on response sets and test design. *Educ. psychol. Measmt.*, 1950, **10**, 3–31.

CRONBACH, L. J. *Essentials of psychological testing.* (2nd ed.) New York: Harper, 1960.

DINGMAN, H. F. The relation between coefficients of correlation and difficulty factors. *Brit. J. statist. Psychol.*, 1958, **9**, 13–18.

EDWARDS, A. L. *The social desirability variable in personality assessment and research.* New York: Dryden, 1957.

FRENCH, J. W. *The description of personality measurement in terms of rotated factors.* Princeton: Educational Testing Service, 1953.

HOLZMANN, W. H., & BITTERMAN, M. E. A factorial study of adjustment to stresss. *J. abnorm. soc. Psychol.*, 1956, **52**, 179–185.

MAYESKE, G. Some associations of musical preference dimensions of personality. Unpublished doctoral thesis, University of Illinois, 1961.

MESSICK, S., & JACKSON, D. N. Acquiescence and the factorial interpretation of the MMPI. *Psychol. Bull.*, 1961, in press.

ROBERTS, A. O. H. "Artifactor" analysis: Some theoretical background and practical demonstrations. *J. Nat. Inst. Personnel Res., Johannesburg*, 1959, **7**, 168–188.

THURSTONE, L. L., *Primary mental abilities.* Chicago: Univer. Chicago Press, 1938.

TOLLEFSON, D. Response to humor in relation to other measures of personality. Unpublished doctoral thesis, University of Illinois, 1961.

WHERRY, R. J., & GAYLORD, R. H. Factor pattern of test items and test as a function of the correlation coefficient: Content, difficulty, and constant error factors. *Psychometrika*, 1944, **9**, 237–244.

(Received July 20, 1960)

Psychological Bulletin
1961, Vol. 58, No. 2, 175

COMMENTS ON CATTELL'S PAPER ON "PERTURBATIONS" IN PERSONALITY STRUCTURE RESEARCH

WESLEY C. BECKER
University of Illinois

Cattell's reply to my earlier paper (Becker, 1960) questioning the validity of published statements of a one-to-one matching between L-data and Q-data factors concedes the inaccuracy of those statements (see the first point in his summary). However, in the process of developing a defense for his basic *theoretical* position, Cattell has distorted the nature of my arguments to the point that a further brief clarification is needed.

Cattell states several times in his paper (Cattell, 1961) that since the evidence did not support his theory, I concluded that the evidence disproved his theory. In rebuttal I need only quote two sentences from my earlier paper.

> It is apparent that the present evidence does not support the claim for "secure linkage" of BR and Q factors. This does not necessarily imply that future research using more reliable and factor pure measures may not still prove Cattell's proposition to be correct (p. 208).

My critique was based on a question of fact, not of theory. Cattell has conceded this question of fact, as he must, but then he sets up for attack a question of theory which I did not raise. I did go on to indicate on *logical* grounds why I felt complete confirmation of his theory was exceedingly unlikely, and I see nothing in his present paper to change this opinion. The demonstration of a few "matchings" in the extraversion area, where on psychological grounds one would most expect self-perceptions and behavior ratings to overlap, can hardly be accepted as firm evidence for his general theoretical position.

REFERENCES

Becker, W. C. The matching of behavior rating and questionnaire personality factors. *Psychol. Bull.*, 1960, **57**, 201–212.

Cattell, R. B. Theory of situational, instrument, second order, and refraction factors in personality structure research. *Psychol. Bull.*, 1961, **58**, 160–174.

(Received December 23, 1960)

Psychological Bulletin
1961, Vol. 58, No. 2, 176

CATTELL REPLIES TO BECKER'S "COMMENTS"

RAYMOND B. CATTELL
University of Illinois

In his additional comments, Becker expressly concedes that my theory has not been disproved. It is still odd that he objects to my saying that he considered the theory untrue, since he again says that it is "exceedingly unlikely," and, to a scientist, "true" and "untrue" mean "highly probable" or "highly improbable,"—at least, since the time of Victorian physics.

The positive conceptual and experimental contributions of my paper appearing since his comments, he either misses or ignores, since they show: (*a*) that it was impossible for him to reach any intelligible conclusion on the theory without recognizing and developing the necessary corrections for attenuation and perturbation, and (*b*) that the facts which he says I must and do recognize are those chosen by Becker from experiments with older techniques. Science moves on, and the new facts which I present from technically more advanced designs show that *the same factor simultaneously loads on the hypothesized markers for both the rating and the questionnaire factors.* His statement that I concede his facts is therefore ambiguous.

(Received December 26, 1960)

Vol. 58, No. 3 May 1961

Psychological Bulletin

METHODOLOGY AND RESEARCH ON THE PROGNOSTIC USE OF PSYCHOLOGICAL TESTS

SAMUEL C. FULKERSON and JOHN R. BARRY[1]
Western Psychiatric Institute and Clinic, University of Pittsburgh

There has not been a general review of the use of psychological tests in prognosis since Windle's review in 1952. At that time Windle concluded that, (*a*) it appeared to be some characteristic of the patient rather than the therapy given which determined the outcome of mental illness; (*b*) most studies in the area were difficult to interpret due to inadequate specification of one or more of the following: the sample characteristics, the treatment schedule, the criteria of improvement, and the degree of control imposed on variables influencing outcome; (*c*) the necessary step of cross-validation was usually omitted; and (*d*) personality tests, including the projective tests, had shown little promise in predicting outcome.

The purpose of the present article is to bring the review of the research on the prognostic use of tests up-to-date and to deal with some related methodological issues. The scope and organization will depart from that used by Windle. Firstly, the present review covers a wider range of criteria. Windle considered primarily the problem of predicting improvement. However there seems to be a complex of criteria which are closely related, logically and in practice, and so articles have been included dealing with a variety of criteria other than improvement. Secondly, the organization will differ from Windle's. He centered his review around individual tests, taking each test in turn and citing all prognostic studies where it had been used. The present paper is organized around the predictive problem rather than the individual test, since in practice the clinician wants to know how to come to a decision about a patient rather than what can be done with a given test. It is hoped that this emphasis will help to point up which questions are involved in the area of prognosis, and the relative attention each has received in research. And finally, the emphasis on decisions reflects an interest in decision theory (Luce & Raiffa, 1957), which has recently been suggested (Cronbach & Gleser, 1957) as a promising frame of reference from which to regard psychodiagnostic testing.

Windle included studies from as early as 1926 through 1951. The present review mainly covers the period from 1952 through June 1959. The coverage is more complete for those sections dealing explicitly with the prognostic use of tests than for the sections on methodological problems. Only the major psychological and psychiatric journals have been reviewed exhaustively.

[1] Acknowledgment is due the criticism and additions of Charles Windle and Joseph Zubin.

METHODOLOGY

The methodological difficulties in research on prognosis concern the researcher's decisions as to what samples he will use, what selection instruments he will apply to the sample, and what criteria seem most appropriate.

Sample Attributes

One of the primary methodological difficulties has been the definition of the sample. Psychiatric diagnosis appears to have been the predominant basis of sample definition in spite of the known unreliability of these categories. Attributes of the sample such as age, education, sex, or socioeconomic status are usually listed. However, little attention has been paid in most of the studies reviewed to achieving homogeneous samples or subsamples. Some investigators have worked with only one diagnostic group, mainly schizophrenics. Since schizophrenia is a diagnosis given to over 50% of unspecified functional psychotic disorders, the difference between the results from such studies and those using psychotics sampled at random are hard to determine.

The need for homogeneous samples is clearly pointed up by a consideration of the question of base rates. Meehl and Rosen (1955) said "a psychometric device, to be efficient, must make possible a greater number of correct decisions than could be made in terms of the base rate alone" (p. 194). Studies of base rate as a function of diagnostic category (Langfeldt, 1956; Pascal, Swensen, Feldman, Cole, & Bayard, 1953; Rennie, 1953) indicate wide fluctuation in outcome between categories. Examination of these base rates indicates that a sample of psychotic patients with a preponderance of manic-depressives would have a higher base rate of improvement (approximately 68%) than a sample consisting of schizophrenics (approximately 50%). A predictor, even though its actual validity was zero, could do much better predicting outcome in the first sample than in the second if the cutting point of the predictor was adjusted to take advantage of the percentage of improvement. The optimal chance percentage of correct prediction in the first sample (achieved by calling everyone improved) would be 68%, which is equal to the base rate. With a sample of any size this would differ significantly from 50%, the value likely to be designated as chance if one did not know the base rate. And if the relative effectiveness of a predictor in the two samples were tested it could appear, spuriously, that the predictor was 68% correct within one sample but only gave 50% correct prediction in the other. Thus, prognostic research designs which compare the results in the experimental group against statistical chance, or which compare two small groups that are not sufficiently matched on variables related to base rates, cannot result in useful information. Since effective handling of the problem of sample homogeneity is uncommon in the prognosis studies reviewed by Windle and ourselves, the generality of findings is low, or at best difficult to determine.

It has been assumed that homogeneous sampling represents an effective way of solving the problem of sample definition. However there is one danger. If the basis on which the homogeneity is established is highly related to the criterion variable, the variability of the criterion will be restricted. This can of course obscure a relationship that might exist between a predictor and criterion. It has been tacitly assumed that adequate randomization is not easily achieved in prognosis research, considering the usual sample size and the biases of the clinical populations from which they are drawn; otherwise random sampling would be an efficient way to select, and thus operationally define, the sample.

Another difficulty with diagnosis as a basis for the definition of the samples is that it represents clinical judgments which are based upon an often uncertain weighting of situational and response variables. For instance, the diagnosis of depressive reaction typically requires a differentiation as to whether the affective response reflects anxiety or depression, and a decision concerning the degree to which the affective response is related to a currently stressful situation. Clearly clinicians can vary as to the relative emphasis they place on these variables; and, as several studies (Glass, Ryan, Lubin, Reddy, & Tucker, 1956; Gleser, Haddock, Starr, & Ulett, 1954) have shown, they do vary in their weighting procedures. Therefore, despite the convenience of using diagnosis as a sample-defining operation, it is weak in that the researcher loses some control over the basic stimulus and response elements upon which the judgment is based. Studies of the effects of these elements on test validity and on the efficiency of cutting scores are called for; this kind of research, frequent in personnel psychology, seldom appears in the clinical journals.

Tests

Here the primary difficulty has been to define that universe of test behavior related to outcome. The majority of studies cited in Windle's earlier review used standard tests, e.g., Rorschach, TAT, MMPI. It is likely that these tests tap only a small part of the response spectrum. With the welter available, it is still far from clear how many separate functions they sample, and no definitive taxonomy of tests exists. Zubin and his co-workers (Burdock, Sutton, & Zubin, 1958; Burdock & Zubin, 1956; Zubin, 1958, 1959) have proposed five broad categories of activity to which test behavior can be assigned: physiological, sensory, perceptual, psychomotor, and conceptual. Each of these five categories has been further subdivided into classes of stimuli and responses. For the most part, the prognostic measures which Zubin has selected to use within each category are simpler than such tests as the Rorschach, in the sense that they present fewer stimulus dimensions and require less elaborate and lengthy responses. Such systems of categorization may indicate a range of tests available, but within each category there is a degree of complexity which at this time is largely unknown. However, factor analyses have been carried out in the areas of perception (Thurstone, 1944), psychomotor tests (Fleishman & Hempel, 1954, 1956; Hempel & Fleishman, 1955; Seashore, Buxton, & McCollum, 1940), and cognition (Guilford, 1956, 1959). Such analyses afford at least a partial basis for rational test selection.

Since a number of studies (e.g., Conrad, 1954) indicate that severity of mental illness is a significant prognostic variable, researchers looking for simple tests for prognostic studies may find it of value to consider studies in the area of differential diagnosis. H. E. King (1954) was able to differentiate between chronic schizophrenics, subacute behavior disorders, and normals using psychomotor tasks; and Eysenck, Granger, and Brengelmann (1957), with groups similar to those used by King, found a large number of both simple and complex perceptual tests which discriminated between their groups. Rabin and G. F. King (1958) reviewed studies dealing exclusively with schizophrenia, and concluded that "relatively high discriminatory power . . . has been obtained with simple experimental tasks. In many cases it has been as good as or better than that found with more complex tasks" (p. 253).

Criteria

There are three broad aspects of prognosis in mental illness: duration, course, and outcome. Studies predicting duration have used criteria such as length of hospital stay, the amount of time spent on the admitting or disturbed ward before transfer to a less disturbed ward (Gordon, Lindley, & May, 1957), and length of treatment.

Criteria involving the course of illness include measures of termination and relapse. In inpatient settings premature termination has been defined as leaving the hospital against medical advice; in outpatient settings it has been variously defined as not appearing for the initial interview after making an appointment, dropping out of therapy before some stipulated minimal number of contacts, or dropping out of therapy against the wishes of the therapist.

All criteria of improvement have been classified in this paper as measures of outcome. It could be argued that change over time is a measure of the course of illness, but this category has been reserved for specific qualitative aspects of change. Current criteria of improvement present the same difficulties in definition, and for this reason it is convenient to deal with them together, and to separate them from termination and relapse criteria. Because there is no universally agreed upon definition of the term "mental illness" (Jahoda, 1958; Scott, 1958), there has been a concomitant lack of clarity about how to measure its alleviation. Three sources of improvement criteria are common: (*a*) ratings of improvement made by the therapist, the patient, or other persons in contact with the patient such as relatives, professional staff other than therapist, or even fellow patients; (*b*) changes in objective measures of functioning such as physiological changes, or improvement in psychological test performance; and (*c*) follow-up data of a behavioral nature, such as whether patient is able to get and hold a job, to get or remain married, or, in whatever way, to resume a minimally independent social existence.

There have been several attempts to systematize these various outcome measures. An early breakdown of the separate areas of behavior which should be evaluated was made by Knight (1941). He suggested that therapists look for change in these five areas of adjustment: the disabling symptoms or problems, the interpersonal relations, the sexual adjustment, the productivity (i.e., the ability to work effectively and to utilize available energy), and the ability to handle stress. In Zubin's classification of tests, the ability to handle stress is viewed as a general parameter which might apply to the other four areas.

Barron (1953b) listed five similar criteria of improvement: (*a*) the patient feels better—indicated by introspective comments by the patient; (*b*) the patient relates better to others —requiring a follow-up at work, school, or home, and often based on reports of members of the patient's social group; (*c*) the patient's symptoms clear up—as measured by psychiatric ratings of improvement at discharge, as well as indirectly by measures of duration, e.g., length of hospital stay and speed of transfer to minimum security wards; (*d*) the patient makes decisions in a health-tending direction; and (*e*) the patient's verbal behavior shows increased "insight."

A few other criteria have occasionally been proposed to supplement these. Winder (1957) has suggested changes in the adjustment of children of the patient, and Morse (1953) has proposed accessibility to psychotherapy. Reznikoff and Toomey (1959) list in detail a variety of attempts to provide a taxonomy of outcome criteria.

There are measurement problems in all of these approaches. Scott (1958) has pointed out several conceptual and methodological difficulties in the various definitions of mental health. His discussion can be applied to Barron's criteria of improvement in mental health: (*a*) apparent change in subjective feelings or symptomatology can be a function of change in environmental conditions or can be distorted by defense mechanisms; (*b*) difficulties in social relationships can be a function of the differing requirements of socioeconomic and cultural systems, and can change as the patient changes his community or his contacts in the community; (*c*) there can be disagreement over which is a health-tending direction, since value systems are frequently involved; and (*d*) changes in insight may be a function of the degree to which the patient is willing to conform to the theory and values of the therapist. It should be noted that these points need not be regarded as criticisms of the definitions. If, for instance, changes in subjective feelings are considered important in their own right, then changes in feelings, whether due to environment or defense mechanisms, are still of interest. However, when used as criteria, such changes are meant to reflect specific intra-individual changes that are independent of environmental or irrelevant personal factors. Despite this, most of the research in prognosis seems designed to demonstrate only that characteristics of the patient exist which relate to outcome, without controlling sufficiently for the above mentioned environmental and personal factors.

On a less general level, Parloff, Kelman, and Frank (1954) have listed several common sources of ambiguity in improvement criteria: (*a*) improvement is often treated as a unitary concept, but this may be erroneous; (*b*) the emphasis of the rater can interact with aspects of the treatment—for instance, symptoms typically disappear before insight occurs, so that a rater who requires signs of insight before he gives a rating of improvement will judge fewer patients to be improved than one who accepts symptom alleviation as improvement; and (*c*) improvement is likely to be overestimated, since patients fluctuate in behavior, and at any given time signs of improvement in one or more specific areas are likely to be present and thus overvalued by a judge being asked to make a global, subjective rating. Pascal and Zax (1956) criticize the usual gross "improved-unimproved" criterion on the grounds that it is not sufficiently tailored to the specific desired changes of the patient. They reject all nonbehavioral criteria of improvement, and essentially appear to feel that symptom-change should be the primary criterion of improvement.

It would be valuable to know the factorial structure of the above course, duration, and outcome measures. While no study was found which attempted to do this, several reported intercorrelations between two or more prognostic criteria. These will be described separately for the kinds of criteria involved.

Correlations between outcome measures. Kelman and Parloff (1957) intercorrelated a number of measures, including ratings of comfort and self-awareness made by the patient, and social effectiveness ratings made by persons close to the patient as well as professional observers. The change in rating from pretherapy to 20 weeks after the initiation of therapy was determined. Only 1 of 21 intercorrelations between these measures of change was found to be significant. However, the correlations were based on an *N* of only 15, and the period of time was perhaps too short to expect more than minimal changes.

Storrow (1959, 1960) compared ratings of improvement made by therapists, patients, relatives of the patient, and a psychiatrist who had access only to abstracted material. Two related rating clusters were found: the patient's self-rating, the relative's rating, and the rating made by inexperienced therapists (third year medical students) formed one cluster; with the experienced therapist and the nontherapist psychiatrist forming the other. The correlations within clusters ranged from .61 to .79; between clusters, .32 to .57. These two clusters seemed to reflect primarily a dichotomy between patient and experienced therapist, since the relatives, and apparently the inexperienced therapists, gained their impression from hearing the patient's views of his progress, while the nontherapist psychiatrist obtained his knowledge from the file written by the therapist. Storrow had the ratings made separately for each of Knight's (1941) five areas, and the average intercorrelation between areas was approximately .60. Ellsworth and Clayton (1959) found that a measure of ward adjustment at discharge correlated significantly (.47) with a 3-month follow-up rating of community adjustment. However, amount of psychopathology at discharge had no relationship to the follow-up criterion. Their finding can be compared with the intercorrelation of .57 reported between two simultaneous ratings of adjustment made on different scales (Stilson, Mason, Gynther, & Gertz, 1958).

Patient expressions of positive and negative feelings have been used as evidence of improvement (see Auld & Murray, 1955, for a review of these measures). Barry (1950) found low but significant correlations between these so-called internal or feeling criteria and global judgments of improvement in adjustment. Rogers and Dymond (1954) have found that changes in patient self-ratings on *Q* sorts correlated with ratings and other criteria of improvement. In an analogous group research program Snyder (1953) reported that self-rating changes correlated significantly with judgments of improvement. The same results have been reported by Kalis and Bennet (1957). Taylor (1955) found that self-ratings (*Q* sorts) tend to become increasingly positive simply with the passing of time. This suggests that it is imperative to control for time in treatment in order to evaluate the actual extent of the relationship between self-ratings and other improvement criteria.

Correlations between duration and outcome. Ullman (1957) reported that a measure of length of hospital stay correlated .36 ($N=72$) with a measure of adequacy of interpersonal relationships (Palo Alto Group Therapy Scale), those rated most adequate after a period of group therapy being the ones with the shortest hospital stay. Pascal et al. (1953) found a correlation of .37 ($N=486$) between length of hospital stay and ratings of improvement made a year after discharge; again, the greater the improvement, the shorter the hospital stay.

A significant positive relationship has been frequently reported (Bailey, Warshaw, & Eichler, 1959; Myers & Auld, 1955; Seeman, 1954; Sullivan, Miller, & Smelser, 1958) in which greater length of psychotherapy in outpatient settings is accompanied by judgments of greater improvement. An interesting exception to this is the phenomenon called the "failure zone."

D. S. Cartwright (1955) found a grossly linear relationship between the number of psychotherapy sessions and success of outcome as noted by the therapists; but the mean success rating dropped sharply for those whose therapy lasted from 13 to 21 interviews. Cartwright was reporting on cases treated by nondirective techniques. Taylor (1956) validated this "failure zone" in a psychoanalytically oriented setting. Standal and van der Veen (1957) obtained the same drop in a counseling center sample. Vosburg (1958), in an examination of treatment charts, found evidence that from the fifteenth to twentieth hour was a period where outpatients tended to be preoccupied with their relationship with the therapist, suggesting that treatment which ended in this period might often be due to a desire on the part of either the patient or therapist to avoid the close, dependent relationship which was developing. Perhaps supplementing this, Ends and Page (1959) reported that the "flight into health" reaction occurred in group psychotherapy uniformly around the fourteenth session.

Correlation between duration and course. Crandall, Zubin, Mettler, and Logan (1954) found a significant relationship between the duration of initial hospitalization and rehospitalization; patients who stayed in the hospital a short time were most likely to still be out of the hospital on 1 to 4 year follow-up.

To summarize these intercorrelations, patient self-ratings and therapist ratings appear to covary to a high degree. Although measures of duration and course of illness have some relationship to improvement ratings, they seem also to tap different sources of variance.

Reliability. The reliability of outcome criteria has received attention; the duration and course measures are objective enough so that their reliability has been taken for granted. Miles, Barrabee, and Finesinger (1951) reported low interjudge but high test-retest intrajudge reliability of global judgments of improvement. Ten cases were rated by four judges on a six-point scale. There was complete agreement for only 20% of the judgments, though no disagreement was by more than two points. Test-retest figures showed 70% to 74% complete agreement between ratings taken 6 to 8 months apart. The ratings were based on structured interview material, and probably represent the lower bounds of interjudge agreement, if it is assumed that ratings made after a long period of observation of the patient would show more stability than ratings made on the minimal information contained in a structured interview. These investigators felt that changes in psychiatric status over time cannot be discriminated any more finely than in terms of three gross classes: unchanged or worse, improved, and markedly improved. Levitt (1957) presented data suggesting that judged improvement rate tends to increase as a function of the number of points on the scale. The greatest discrepancy was due to studies using a two-point "improved-unimproved" scale, where the mean percentage improved was 51. Studies using three- to five-point scales had mean improvement rates of 73% to 76%.

A possible source of unreliability in judgments of improvement lies in the fact that they may confound the amount of change with the absolute level of terminal adjustment. Thus it seems likely that the reason initial severity of illness correlates with improvement (Conrad, 1954) is to some extent due to the fact that those who are high on a measure of adjustment initially will be high on adjustment terminally, though the change may be far from being as dramatic as for patients who are admitted in a state of confusion and disorientation, and discharged without these symptoms. Since each judge can combine amount of change and absolute level as he chooses, in most studies, a lowering of interjudge agreement is to be expected. This may be involved in the much higher interrater reliabilities reported by Morton (1955) than by Miles, Barrabee, and Finesinger (1951). Morton developed seven-point scales of absolute level of adjustment in 12 different areas. After training, the interrater reliability coefficients ranged from .79 to .91 when the ratings were based only on transcriptions of a terminal interview; and the reliability of the improvement score (the difference between ratings of an initial and terminal interview) ranged from .59 to .78.

Tests as criteria. A possible criterion of outcome is performance as measured by tests. The present review uncovered no studies which used changes in test scores as primary prognostic criteria but it remains a reasonable possibility. The primary requisite for this use of tests would be evidence that the tests covary with the changes in patients that go to make up the concept of improvement. A number of studies have been published which tackle this question, and in general they support the assumption of covariation.

Pascal and Zeaman (1951) found that the Bender-Gestalt, color-naming, noun-naming, and serial subtraction, from a larger battery of tests, correlated with the course of progress as judged clinically, for four patients getting electroconvulsive therapy.

Hybl and Stagner (1952) reported a significantly greater decrease in the amount of disruption of performance brought about by a frustration experience, for patients rated by their therapists as improved. The tasks were three psychomotor tests: the Ferguson Form Boards, Digit Symbol from the Wechsler-Bellevue, and the Minnesota Rate of Manipulation Test.

Vinson (1952) administered a mirror drawing test before and during electroshock therapy to 18 inpatients. Change in the mirror drawing score correlated .72 with change in orientation as evaluated by the clinical staff.

Several studies (Hozier, 1959; Wechsler, 1958) indicate that as psychotic patients improve there is a decrease in variability of both the quality and the quantity of test performance.

The MMPI has been used in a number of studies of change: several studies (Carp, 1950; Feldman, 1951; Schofield, 1950, 1953) have reported that hospitalized patients treated with somatic therapies show an average drop on all of the MMPI scales of from 8 to 13 *T*-scale points. The acutely ill changed more than the chronically ill, and the affective disorders showed a greater change than the schizophrenics. Feldman (1951, 1952) found that improved patients' MMPI profiles dropped more than unimproved patients' profiles, and that the averaged profiles of these two groups showed greater differences after therapy than before. Work with predominantly psychoneurotic samples (Barron & Leary, 1955; Kaufman, 1950; Schofield, 1950) has indicated a larger drop on most scales for improved patients than for those rated unimproved. Changes taken without regard to sign (decreases as well as increases) were significantly greater in an individually treated group than in a group treated by group-therapy methods (Barron & Leary, 1955; Leary & Harvey, 1956).

Harris (1959) has summarized such MMPI studies to date as follows:

> scores on the MMPI show little change in normals and in untreated psychiatric patients over extended periods of time; somatic therapy, which is known to be effective at least in readying patients for discharge from the hospital, is accompanied by sizeable drops in test scores; patients in psychotherapy show smaller changes, perhaps not much larger than those produced by the passage of time alone; and the magnitude of change in test scores is related to clinical estimates of improvement (p. 519). (Quoted by permission of National Academy of Sciences–National Research Council)

Extraneous effects in test-retest comparisons need to be kept in mind, and Windle (1954) has reviewed these in reference to questionnaires. He presents evidence for a general tendency toward less deviant answers on retest, irrespective of external factors. This tendency is less, the greater the time period between test administrations. But even taking these artifactual sources of error into account, there appears to be evidence that a variety of test responses change in a manner consistent with therapist judgments of change in mental health.

Research in Prognosis

This section is organized around the three elements that seem most prominent in any treatment: the treatment itself, the person administering the treatment, and the patient who receives the treatment. Duration, course, or outcome of illness can potentially be affected by any one of these. The practical need to determine the prognosis of a patient implies that some selection is possible concerning the most appropriate treatment for that patient, or the most appropriate patient for a given treatment. Thus in the headings below we use the terms: treatment selection, therapist selection, and patient selection.

Treatment Selection

Ideally, the basic problem in prognosis is the assignment of patients to treatments in such a way as to maximize the total ratio of improved to unimproved patients. In decision theory terms, the prognostic judgment is a case of decision-making under conditions of certainty, which implies that the relationships between treatments and effects or outcomes are known. However, it has not been demonstrated that different treatments have different effects. To quote an authority,

> One is reluctantly forced to admit that we simply do not possess the factual knowledge as of 1957 which permits us to say that we have any treatment procedure in psychiatry which promises a better outlook for a particular illness than does nature left to her own devices (Hastings, 1958, p. 1057). (Quoted by permission of the *American Journal of Psychiatry*)

Several attempts have been made to survey the literature on treatment effects, all of them hampered by the difficulties in comparing studies with different diagnostic groups, and different criteria for improvement. Eysenck (1952) selected 24 studies on the effect of psychotherapy with psychoneurotics, and concluded that these relatively homogeneous studies did not offer any evidence that improvement rate for those receiving psychotherapy was greater than for those getting only custodial care. Methodological weaknesses in his survey were pointed out by Rosenzweig (1954) and DeCharms, Levy, and Wertheimer (1954).

Levitt (1957) surveyed 30 articles evaluating psychotherapy with children. He compared the improvement rate on discharge and follow-up for treated cases with that reported for children accepted for therapy who never appeared for a first interview. The results were similar to those found by Eysenck, and did not demonstrate any facilitation of recovery due to psychotherapy.

Appel, Myers, and Scheflen (1953) summarized the results of studies which met a list of what they felt were minimal standards. They broke down the findings separately for schizophrenic, affective, and psychoneurotic disorders. Their survey indicated that none of the treatments studied—insulin coma, electroconvulsive shock, electronarcosis, lobotomy, or psychotherapy—gave recovery rates significantly greater than that reported for groups receiving only routine hospital care, in any of the three disorder categories. A more recent review by Staudt and Zubin (1957) covering the somatotherapies indicated that insulin and electroconvulsive shock temporarily increase the improvement rate, but after 3 years the increase has dissipated. This conclusion would seem to fly in the face of the fact that most of the studies reviewed by Staudt and Zubin reported significantly greater recovery for the treated group than for the control group at all periods of follow-up. However, the groups were equally different before treatment was begun; in most instances the control groups "seem to be highly selected and loaded with patients of apparently poor prognosis. Their improvement rates fall far short of the 'spontaneous improvement rates' " (Zubin, 1959, p. 344). This bias in selection of control groups is also likely to be operating in studies of psychotherapy unless matching procedures are possible, since there seems to be a feeling in many clinics that ethical considerations make it mandatory that patients who appear treatable be given treatment as quickly as possible.

Kramer and Greenhouse (1959) discuss a point which bears directly on the adequacy of studies in this area. They show the statistical implications of the common sense notion that the less dramatic the effect one is looking for, the larger the sample necessary to show that it is significant. Their tables indicate that if one is interested in identifying in the experimental group as slight an improvement as 5% over the control group (at the .05 level of significance) for base rate improvement of 40% (which is close to that found in schizophrenia) it would take at least 569 cases in each group. For a base rate improvement of 70% (typical of the psychoneurotic) 472 cases per group would be needed to demonstrate a 5% increase under ideal conditions. These estimates further assume perfect reliability of the improvement criterion. Kramer and Greenhouse point out that very few states have a large enough population of mentally ill to do a study with a sample sufficient to detect slight but significant effects. Thus all the studies on the effect of treatment using small samples implicitly assume no interest in detecting anything less than extremely large differences. This is why it has been emphasized that treatment effects seem to be negligible relative to other variables in determining outcome; in view of the size of samples for research in this area it would not be fair to say that slight treatment effects may not exist.

How do patients regard psychotherapy? Stotsky (1956a) found that only 10% of a VA sample mentioned psychotherapy when asked to list any treatments which helped them. If asked directly whether they felt psychotherapy was the most important part of their treatment, over 50% said yes. These patients came predominantly from a lower socioeconomic class which, as will be discussed later, would bias the results in the direction of more negative answers.

Two final points can be made. It first should be said that clear-cut effects of psychotherapy seem to have been demonstrated using the patient's verbal behavior, rather than judgments of improvement, as the criterion measure (Rogers & Dymond, 1954; Rosenthal, 1955).

Secondly, it might be pointed out that the inconclusive state of affairs regarding the effects of treatment is not necessarily discouraging from the restricted point of view of the researcher. If treatment effects are currently less important than effects due to other sources of variance, then the researcher can ignore treatment differences in his samples and in the formulation of his hypotheses, thus considerably simplifying the research design.

Therapist Selection

A special aspect of treatment selection is the question of what kind of therapist does best with what kind of patient in psychotherapy. In the years surveyed in this review the pertinent articles in this area dealt with such therapist variables as sex, vocational interests, professional affiliation, and experience.

Irrespective of cause, are there differences between therapists as to treatment results? Imber, Frank, Nash, Stone, and Gliedman (1957) compared three therapists, each of whom worked with 18 patients. No significant differences were found between therapists, against a criterion of ratings of improvement in social effectiveness. Sullivan, Miller, and Smelser (1958) found neither sex, experience, nor profession (psychiatrist, psychologist, or social worker) to be related to either length of stay in therapy or to ratings of improvement. Hiler (1958a) reported significant differences in number of responses on the Rorschach between six groups of patients (14 per group), each group subsequently treated by a different therapist. He interpreted this as indicating that the therapists differed in their ability to keep unproductive patients in therapy. Stieper and Wiener (1959) found significant differences between therapists as to the length of time they kept patients in therapy. The differences seemed to be related to personality variables in the therapist, such as having high goals concerning very sick patients, and needing to feel appreciated. They took a negative view toward this minority of therapists who keep patients in therapy for long periods:

> It seems to us likely that psychotherapeutic practice today contains self-defeating concepts which may not only be hampering to the success of treatment, but potentially harmful to its clients (p. 241).

Betz and Whitehorn (1956) found differences in treatment between therapists who had a cumulatively high improvement rate with schizophrenics and therapists with a low improvement rate. The successful therapists were more active, emphasized utilization of assets, understood the meaning of the patient's behavior, and engendered more trust and confidence. They also differed from unsuccessful therapists in their scores on the Strong Vocational Interest Test.

Myers and Auld (1955) found that the experienced staff in an out-patient clinic had fewer patients quit against the therapist's wishes, and more patients who improved, than the residents in the same clinic. Katz and Solomon (1958) concluded that in their sample the less experienced therapists tended to lose more patients, but if the patient continued treatment, the improvement rate was as high as for the more experienced therapists. Strupp (1958) had 134 residents and psychiatrists respond to a sound film of an initial interview. He interpreted his data as showing two types of therapists. Type I was positive in his feelings toward the patient, optimistic about prognosis, and permissive and passive in therapy—and relatively inexperienced. Type II was more experienced, was negative toward the patient, pessimistic about prognosis, and active in therapy (giving orders and advice, and venting his irritations). Strupp quotes Kubie (1956) on reasons for this increasing pessimism: Kubie mentions his disappointment, saying it is one shared by other psychoanalysts, to find that with increasing experience he did not seem to have increasing success.

Several studies (Katz, Lorr, & Rubinstein, 1958; Sullivan et al., 1958) have reported that the more experienced the therapists, the larger the percentage of cases rated by him as improved; and the less severe the illness, the greater the likelihood of a patient's having an experienced therapist. Clearly, it is advisable to control for severity of illness in research on therapy. Differences in socioeconomic level also appear to interact with experience. Schaffer and Myers (1954) studied all cases accepted for treatment in an out-patient clinic during 1 year and found that

> the higher a patient's social class position . . . in the community, the greater were his chances of being accepted for psychotherapy, of being assigned to a relatively experienced therapist occupying a high status within the clinic, and of maintaining contact with the clinic (p. 88). (Quoted by permission of *Psychiatry*)

It is apparently also likely (Winder & Hersko, 1955) that the higher the social position, the higher the likelihood that the therapist will decide on analytic rather than supportive procedures.

Since the above studies did not control for these contaminating factors, it must be concluded that demonstration of between-therapist effects on outcome has not been conclusively obtained. This is not particularly surprising, in view of the fact that therapist selection is just a special case of treatment selection. Again, though, it can be said that effects can probably be shown, against other than improvement criteria. For instance, Rosenthal (1955) found that the amount of benefit a client said he obtained from therapy correlated .68 with the degree of shift in moral values toward those held by the therapist, if the values had been talked about during psychotherapy. This change would appear to be related to those obtained in laboratory studies on verbal conditioning (Krasner, 1958).

Patient Selection: Outcome Criteria

We turn now to the question of the relationship of intra-individual variables to prognostic criteria. The studies will be grouped along two dimensions. They will be considered according to the kind of criterion used—outcome, duration, or course—and further broken down, where possible, in terms of the type of test used—projective, questionnaire, or performance (including cognitive tests).

Nontest indicators. Before turning to the research using tests as predictors of outcome, it is of interest to survey briefly what has been found using nontest variables. Huston and Pepernik (1958) reviewed prognostic variables in schizophrenia, and presented evidence that only these variables had been firmly established as going with favorable outcome: acute onset, short duration of illness prior to hospitalization, a precipitating stress, and the absence of flat or inappropriate affect. A series of studies under the direction of Pascal investigated the interrelationships of these variables within a sample of varied psychotics. It was found that acute onset (Swensen & Pascal, 1954b) and aggression directed toward oneself (Feldman, Pascal, & Swensen, 1954) related significantly to favorable outcome when other prognostic variables were controlled. However, precipitating stress (Cole, Swensen, & Pascal, 1954), affective expression (Bayard & Pascal, 1954), and duration of illness (Swensen & Pascal, 1954a) did not relate to outcome in their sample when the effect of other prognostic variables was held constant. The generality of their findings is not clear, since their method of balancing groups for control purposes led to their using only a small portion of the total sample, thus allowing for the possible introduction of unknown biases.

Eskey, Friedman, and Friedman (1957) could not find support for the notion that disorientation relates to duration of illness; however, they restricted their sample on the criterion variable by not using patients who were unimproved at discharge. Several studies (Eskey & Friedman, 1958; Phillips, 1953) indicate that intact cognitive processes and a mature premorbid social and sexual life go with favorable outcome. Zubin (1959) presents the results to date of an uncompleted survey of prognostic indicators for schizophrenia, which suggests that the variables defining reactive schizophrenia go with favorable prognosis, and those defining process schizophrenia go with unfavorable prognosis. He presents a valuable count of articles supporting or negating the postulated relationship for almost every if not every prognostic indicator that has been investigated. There have been several attempts to combine these variables into a scale. Thorne (1952) intuitively combined five into a quantified prognostic scale. More recently Lindemann, Fairweather, Stone, and Smith (1959) have developed a somewhat similar scale and cross-validated it against a criterion of duration of hospital stay. An eight-point scale (Schofield, Hathaway, Hastings, & Bell, 1954) developed to predict a follow-up criterion of adjustment in schizophrenia could not be cross-validated by Stone (1959). Becker and McFarland (1955) developed and cross-validated a 16-item scale against a criterion of improvement in a lobotomized sample.

The above studies have dealt with psychotics, or samples predominantly psychotic. Miles, Barrabee, and Finesinger (1951) reported that in a hospitalized psychoneurotic sample, age of onset, duration of illness prior to hospitalization, and a number of symptoms were unrelated to outcome. Patients with symptoms associated with autonomic discharge were most likely to remit. Rosenbaum, Friedlander, and Kaplan (1956), studying an outpatient sample, found improvement occurred in patients with good premorbid history whose environment offered many supports; and improvement was mainly in marital and work adjustment. Comparison of results on inpatient and outpatient samples suggests some reason for dealing separately with psychotics and psychoneurotics in prognosis research.

An important question is how well the clinician, using these nontest indices, can do in predicting outcome. Clow (1953) obtained a majority opinion of prognosis at the staff conference which was held 2 months after admission on each of 100 female schizophrenics. The prognoses were 73% correct in predicting a dichotomous improved-unimproved criterion obtained at discharge. More studies of this kind would be helpful in evaluating the practical usefulness of adding tests to current prognostic procedures.

Projective tests. Several Rorschach studies have used a configurational score, the Prognostic Rating Scale (PRS) (Klopfer, Kirkner, Wisham, & Baker, 1951). Kirkner, Wisham, and Giedt (1953) found a correlation of .67 between PRS and improvement ratings obtained by evaluating the terminal closure note, on a sample of 40 receiving psychotherapy. Mindess (1953) obtained a correlation of .66 (*N* of 70) between PRS and a diagnostic criterion running from normal through neurotic to psychotic, obtained 6 months after initiation of psychotherapy. Filmer-Bennett (1952, 1955) did not obtain significant results with either the PRS or global judgments based on the total Rorschach protocol. His criterion was a dichotomous improved-unimproved rating of the degree to which the patient was making a satisfactory social and vocational adjustment a year after discharge from the hospital. Rosalind D. Cartwright (1958) presented a review of several successful studies using the PRS, and described further positive results from her own study. The criterion was ratings of success of psychotherapy made by the counselor after termination of therapy. In an appended discussion of her paper Snyder argued that other tests might do as good a job with much less time needed for testing. Bloom (1956) added an interesting modification to his design. He divided his 46 subjects into two groups, an unproductive group (less than 11 Rorschach responses) and a productive group (11 or more responses). The PRS differentiated a dichotomous criterion of outcome of psychotherapy significantly in the productive group, but not in the unproductive. He further assessed 11 other scores, and found none which were either significant or nonsignificant for the total sample as a whole; all discriminated significantly in one or the other of his groups—four for the productive group, and seven for the unproductive. His results suggest the operation of an interaction similar to the one Zubin and co-workers (see below) have reported between chronicity and outcome, and deserve further investigation.

Rogers and Hammond (1953) and Roberts (1954), both working with VA outpatients, tried a sign approach on the Rorschach with negative results. Dana (1954) hypothesized that Card IV, assumed to be most likely to pick up attitudes to authority, would give responses related to improvement in psychotherapy, if the authority relationship was crucial to outcome. The responses were placed in three categories—"adequate," "inadequate," "negative"—and there was a significant tendency for those with "adequate" response to improve, and those with "inadequate" responses to remain unimproved. Hammer (1953) felt that his review of the literature suggested that those patients whose Rorschach protocols look sicker than their H-T-P protocols have a good prognosis, while a poor prognosis is associated with giving more negative feelings on the H-T-P than on the Rorschach.

Ullman (1957) found two highly related measures—clinical judgments of TAT protocols and a social perceptions test—to be correlated significantly with two criteria of improvement: the Palo Alto Group Therapy Scale and hospital status after 6 months (hospitalized vs. discharged). S. Rosenberg (1954) developed and cross-validated eight prognostic signs based on the Wechsler-Bellevue, Sentence Completion, and on the Rorschach. Grauer (1953) found more Rorschach indices of anxiety in an improved group of schizophrenics than in an unimproved. Organic signs did not discriminate. The welter of signs which these studies find related to improvement shows no clear pattern. Obviously most of these positive findings with projective techniques should be further validated before they can be accepted as more than promising leads.

Questionnaires. Barron (1953b) reported lower pretherapy MMPI and Ethnocentrism scores for an improved outpatient group than for an unimproved group. The criterion was judgments of change in psychotherapy made by professionals who had not been involved in the treatment. At least some of these relationships were due to differences in IQ between the groups. Rosen (1954) was not able to verify Barron's finding with the E Scale. Barron developed a special ego strength scale from the MMPI (Barron, 1953a), which he successfully cross-validated against improvement criteria in three disparate samples. Wirt (1955, 1956) found the ego strength scale significantly discriminated an unimproved from a greatly improved group, the groups being extremes drawn from a hospitalized sample receiving psychotherapy. The scale did a better job of discrimination than experienced clinicians who based their judgments on the total MMPI profile.

Feldman (1951, 1952, 1958) explored the validity of the MMPI for the prediction of outcome after electroshock therapy. He found that items dealing with hostility and interpersonal relationships were predictive of outcome, while items dealing with symptomatology reflected the amount of improvement. Pumroy and Kogan (1958) were unable to cross-validate Feldman's prognostic scale in a small VA sample. Dana (1954) also obtained negative results with the MMPI, attempting to predict improvement after electroshock.

Performance tests. Stotsky (1956b) gave vocational aptitude and interest tests to a group of schizophrenics most of whom had been in the hospital for a year or more. The aptitude tests predicted later work success, but the interest tests did not. Swensen and Pascal (1953) reported that the Pascal-Suttell Z score on the Bender-Gestalt test, was significantly lower for a group of inpatients judged to be improved on follow-up a year and a half later, than for those judged unimproved. Landis and Clausen (1955) found efficient performance on critical flicker fusion, reaction time, finger dexterity, auditory acuity threshold, and tapping speed was predictive of improvement in an inpatient sample receiving a variety of treatments. A variability score of palmar sweating (Ellsworth & Clark, 1957) predicted changes in a behavioral adjustment scale concurrent with the administration of tranquilizing drugs. Keehn (1955) took 12 measures from simple cognitive and psychomotor tests that had been shown to discriminate between normals and psychotics, and found only one score that predicted outcome in a group of inpatients receiving insulin coma therapy; he concluded that initial degree of psychoticism was not prognostic of outcome.

Vinson (1952) used a mirror drawing test to predict the prognosis made at discharge—a dichotomous "favorable-unfavorable" prognostic judgment made by the staff. His sample consisted of 18 hospitalized patients who received electroshock therapy. He tested before and during treatment, and the difference between these scores predicted the prognostic criterion at the .02 level of significance.

The most promising findings made in prognosis in the last 10 years have been reports coming out of the Columbia-Greystone project of two interaction effects. The first interaction dealt with the relation of chronicity to outcome. Windle and Hamwi (1953) reported that chronic patients who were discharged after treatment had poorer admission scores on a complex reaction time test than chronic patients who were not discharged. However, for acute patients, those whose illness was of short duration, the reverse was true, namely, poor admission scores were associated with poor outcome. Zubin, Windle, and Hamwi (1953) rechecked data on other tests, using chronic patients from the same study, and found four other tests which gave the same results. An independent validation was provided by Sonder (1955) using different tests. In all of these studies the results were most clear-cut for the chronic group, probably due to the fact that among the acute patients were some who were potentially or actually chronic.

The second interaction emerged from the study by Zubin, Windle, and Hamwi (1953) who found that the chronic patients who did well on conceptual tasks (intelligence, memory, personality tests) but poorly on perceptual tasks (learning and perception tests) had a poorer prognosis than chronic patients who showed conceptual confusion but perceptual clarity. Williams and Machi (1957), also working with the chronic sample from the Columbia-Greystone project, factor analyzed the test data, and found some support for this conceptual-perceptual differentiation. However, this finding is not yet as clearly supported by the evidence as the chronicity-outcome interaction. Zubin and Windle (1954) reviewed a number of independent prognostic studies, and reported that a consideration of the two interaction effects accounted for much of the conflicting findings. In the light of this work, further attempts to investigate these interactions cannot help but be of value.

Patient Selection: Duration Criteria

Projective tests. Stotsky (1952), working only with schizophrenics, compared a group of patients who in a 2-year period had not left the hospital with a group which in the same period of time had been discharged and remained outside for at least 6 months. His hypothesis was that the prognosis would be best for patients with the best pretreatment emotional and intellectual integration. Of 19 Rorschach signs, 5 were significantly cross-validated in a second sample, Also, all of the 19 signs except *R* were found to be in the predicted direction in both samples.

Questionnaires. Grayson and Olinger (1957), in a VA inpatient sample, reported that those who were given early trial visits were able to give improved MMPIs when asked to respond in "the way a typical, well-adjusted person on the outside would do" to a greater extent than those still hospitalized after 3 months. Rapaport (1958) was not able to validate this finding, using a military sample, although the change on most of the scales was in the correct direction. Stieper and Wiener (1959) found a group of VA outpatients who were seen in psychotherapy for an average of 5.3 years had higher pretherapy scores on the MMPI scales, *Hs* and *Hy*, than a group who were discharged after 14 months.

A demographic study (Lindemann et al., 1959) found an index using marital status, diagnosis, degree of incapacity, legal competence, and alcohol intake as variables, was related to length of hospital stay. Ellsworth and Clayton (1959) found a rating scale of psychopathology filled out at admission did not correlate significantly with length of hospital stay, but a behavioral adjustment scale did correlate, patients with the best admission adjustment tending to remain in the hospital the shortest length of time.

Performance tests. Venables and Tizard (1956b) found "short-stay" schizophrenics performed better on a repetitive psychomotor task than did chronic schizophrenics. Reaction time differences (Venables & Tizard, 1956a) occurred on initial testing, but disappeared on retest.

Patient Selection: Course Criteria

Under criteria measuring the course of illness we have placed two broad questions: who will relapse, and who will terminate treatment.

Relapse. The broad question here is one of predicting who will get worse over time. It is of course the reverse of the question of who will improve. However, the prediction of improvement and its opposite may not necessarily be most effectively accomplished with the same test. It can not be assumed that the prediction of relapse or hospitalization can be made from the same tests which predict improvement. This is consistent with the assumption that change of mental status need not be a unitary concept.

Peterson (1954b) used the MMPI, Wechsler-Bellevue, Rorschach, and nontest data to predict who would require admission to the hospital from patients being seen on an outpatient basis in a VA mental hygiene clinic. Considering the base rates, the predictive power of the tests was slight, but the results suggested that the person who gets worse in therapy is single, has been previously hospitalized, is diagnosed psychotic, and has an MMPI profile strongly elevated on the psychotic scales. Using a six-point scale based on signs of psychosis on the MMPI developed by Meehl, Peterson (1954a) was able to achieve 75% correct discrimination. Briggs (1958) was able to cross-validate this scale to a certain extent. He took patients who were already in the hospital when they received the MMPI. On follow-up he found the Peterson score differentiated those who were rehospitalized from those who were not only for patients originally diagnosed psychoneurosis or mixed psychoneurosis. This is consistent with Peterson's finding that in his study similar outpatient diagnoses were most often given to the cases which were later hospitalized.

Schofield and Briggs (1958) related several measures of improvement previous to initial discharge to rehospitalization, the median follow-up period being 5.8 years. Improvement in behavior ratings made by nurses was not related to rehospitalization, but a combination of ratings based on pre- and posttreatment MMPIs and psychiatric evaluations of improvement made at the time of discharge allowed 75% correct prediction for the 66% of cases on which the two ratings agreed. Since knowledge of the base rate alone would allow 66% correct prediction, this was only slightly better than chance.

Cowden, Deabler, and Feamster (1955), using a criterion of whether patient was rehospitalized within 90 days after discharge, reported judgments of change from admission to discharge on Sentence Completion and the H-T-P Test predicted the criterion. An "ego" score obtained from combining the Binet Vocabulary with Cards I, III, and VIII of the Rorschach predicted relapse within a 2-year period for a sample of discharge patients (Orr, Anderson, Martin, & Philpot, 1955), but did not predict discharge for a sample of nondeteriorated admissions. Working with a special group (outpatients considered interminable) Wiener (1959) studied return to psychotherapy over a 6-month period after initial psychotherapy was arbitrarily terminated. In his sample of 48, 37 returned for further therapy within this period. The MMPI did not discriminate returnees from nonreturnees. Months in treatment appeared to be a promising measure, with the returnees having a longer history of psychotherapy.

A study that fits under neither of our two course criteria is one by Rioch and Lubin (1959). They obtained lengthy follow-up data on 93 patients, sufficient to allow an assessment on an 11-point scale of how consistently the patient had moved upward or downward in his social adjustment over several years. Both the Wechsler-Bellevue IQ and a global rating based on the Rorschach correlated significantly with this criterion, mainly due to discrimination at the low end of the scale: all of the patients who deteriorated steadily had low scores on the predictors.

Termination of treatment. The criterion involved in the prediction of length of therapy is more objectively determined than improvement, but there are some difficulties in its determination nonetheless.

One question is how to measure length of therapy. Most studies have used the number of interviews as the measure. Number of weeks in treatment would appear to be an equivalent measure. However, Lorr, Katz, and Rubinstein (1958) found that the number of interviews correlated only .60 with number of weeks in treatment, and they argued that number of interviews is likely to be the less reliable of the two.

Another problem springs from the

research design used in most of the studies of termination. The total sample is usually divided into two groups, terminators and remainers, and test scores are related to this dichotomous criterion. The question becomes one of where to cut the distribution. Terminators have been defined as those remaining less than 4 sessions (Gliedman, Stone, Frank, Nash, & Imber, 1957), less than 10 sessions (Auld & Eron, 1953; Kotkov & Meadows, 1953), or less than 20 sessions (Gibby, Stotsky, Hiler, & Miller, 1954). Gibby et al. (1954) found that those terminating between 5–19 sessions resembled in their test responses those who terminated earlier rather than those continuing on for more than 19 sessions. Our previous discussion of the "failure zone" (Taylor, 1956) suggests that a variety of factors are operating in the first 20 weeks. When these factors have not been controlled, they can influence the findings in termination studies.

A further criticism has been made by Gundlach and Geller (1958) who suggest that termination rate and duration of illness are partly administrative artifacts, and partly a reflection of "the kind of personality problems that the staff are interested in, or skilled at, handling." This criticism can be taken as indirect support for the common practice of defining termination in terms of the distribution of the length of therapy measures, since in any given setting, the median or mean length takes some account of the effects of policy and staff interests.

Research on the prediction of termination by the use of *projective tests* shows a familiar, monotonous pattern: initial positive results with subsequent negative or indeterminate cross-validation. Kotkov and Meadow (1952, 1953) began with 12 formal scores, and validated one of these (*FC/CF*). They applied a formula based on three scores (*FC/CF, R, D%*) to another sample, and *D%* washed out. When these same signs are examined in an earlier study (Rogers, Knauss, & Hammond, 1951), none were significant, and only *R* was in the predicted direction. Auld and Eron (1953) tried a further validation of the Kotkov and Meadow formula, and obtained insignificant results. They found the Wechsler-Bellevue IQ accounted for the one Rorschach variable, *R*, which held up in their sample.

Starting anew, Gibby et al. (1954; Gibby, Stotsky, Miller, & Hiler, 1953) found 9 of 31 Rorschach signs promising. Taking the 9 to a second sample, 3 held up (*R, K, m*) and a predictive formula based on these variables was applied to a further independent sample, and afforded 68% correct prediction. However, knowledge of the base rate would have allowed 60% correct prediction, so the results were not strong enough to be of practical use. In their sample the Kotkov and Meadow formula did no better than chance, and IQ was not related to the criterion. Affleck and Mednick (1959) used an equation based on *R, M*, and *H* to predict who would remain for longer than three interviews. Their equation allowed 71% correct prediction in a validation sample. Their terminators were lower in IQ than the continuers (significant at .06 level). This is consistent with the findings of Auld and Eron (1953).

All of the above Rorschach studies except for Auld and Eron used equivalent VA males being seen on an outpatient basis, so in some respects sample homogeneity was better from study to study than is true of most validation research in this area. Of all the Rorschach signs only *R* seems to have maintained its promise in these studies. More recent work (Gallagher, 1953, 1954; Taulbee, 1958) supports the conclusion that the number of Rorschach responses (*R*) relates to termination. However the Rorschach is probably an unnecessarily cumbersome way of measuring this variable; for instance, Gallagher (1954) found that the number of words used on the Mooney Problem Check List to describe the clients' problems was a better predictor than *R*.

Libo (1957) used a TAT-type test to predict the number of patients who would return the week after the test was administered. For 40 subjects he was able to make a significant prediction based on an "attraction score": the number of references in the stories to a desired move toward the therapist, or of anticipated satisfactions from therapy.

Three studies dealt with the prediction of termination in a tuberculosis hospital. Vernier, Whiting, and Meltzer (1955) were able to differentiate patients who left the hospital against medical advice from those who continued treatment to the end, using the Rorschach and H-T-P tests. The TAT did not discriminate. Moran, Fairweather, and Morton (1956), using a biographical inventory and an attitude questionnaire found that only prehospital life adjustment predicted who would leave the hospital prematurely, with those leaving having a long history of being unable to adjust to their life situations. Calden, Thurston, Stewart, and Vineberg (1955) developed and cross-validated a scale from the MMPI to predict premature discharge.

Taulbee (1958) developed a key based on the MMPI and the Rorschach to predict continuation of outpatient psychotherapy beyond the thirteenth interview. His results, not cross-validated, led him to conclude that those who continue in therapy are less defensive, and more persistent, dependent, anxious, and introspective than the terminators. Sullivan et al. (1958) reported no significant difference between MMPI scores of terminators, and continuers on a VA male sample. Of a number of variables only education and occupation related to the criterion. Conrad (1954) had therapists fill out a check list covering positive mental health, social conformity, and behavior pathology on VA outpatients with differing lengths of stay in psychotherapy. Continuers tended to look least disturbed initially, and to be at the median rather than at either extreme on social conformity.

Rubinstein and Lorr (1956) found differences between extreme groups (patients in psychotherapy for over 6 months vs. patients who had come less than six times and had terminated against the wishes of the therapist), on the authoritarian F Scale, and a vocabulary test. However, a later study (Lorr et al., 1958) which defined termination as having less than 7 weeks of psychotherapy, did not give significant results, though the scales were in the predicted direction. They combined a number of scales in a further attempt, and obtained a significant multiple correlation in a validation sample. However, the scales allowed no better prediction than interviewer's judgment.

A large recent project on termination was carried on at Johns Hopkins University (Frank, Gliedman, Imber, Nash, & Stone, 1957; Gliedman et al., 1957; Imber, Frank, Gliedman, Nash, & Stone, 1956; Imber, Nash, & Stone, 1955; Nash, Frank, Gliedman, Imber, & Stone, 1957). Their prognostic battery included an inventory and a Sway test. Those who stayed in therapy more than three interviews were more suggestible on the Sway test, were more sociable, of higher socio-economic status, and more likely to see treatment as a means of maintaining status in their immediate social environment, than the terminators. When they compared group versus individual psychotherapy they found an interaction between treatment and termination: in group therapy, the terminators were more socially ineffective than the continuers, while the relationship was reversed for those getting individual therapy. This intriguing finding may have been related to an unequal distribution of social levels in the two groups—most of the lower class patients ended up in group psychotherapy, while most of the middle class patients were assigned to individual psychotherapy.

Hiler (1959) studied intial complaints, and concluded that continuers come to a clinic with typical psychoneurotic symptoms—obsessions, phobias, anxiety, depression, poor concentration—while early terminators are more likely to list purely organic symptoms, antisocial acts, or schizoid feelings. His continuers also obtained higher scores on the Wechsler-Bellevue with a subtest pattern characterized by Similarities being higher than Digit Span or Digit Symbol (Hiler, 1958b).

How much overlap is there between predictors of termination and improvement? Sullivan et al. (1958) investigated the relationship of MMPI scores and demographic variables to both improvement and termination criteria. Only occupational level was related significantly to both. Katz et al. (1958) found none of their predictors of length of stay correlated with therapist ratings of improvement. Frank et al. (1957) reported that a past history of social activity and a fluctuating course of illness was associated with continuation and improvement. A short duration of illness was associated with termination as well as improvement. Gallagher (1954) found the Taylor Manifest Anxiety Scale predicted continuation as well as improvement. In general the results suggest little overlap. This is somewhat unexpected, since as was mentioned earlier, there appears to be a positive relationship between criteria of duration of treatment and improvement. The most tenable assumption would seem to be that the variance shared by the two criteria is different from the variance shared by predictor and criterion. Possibly the correlation between criteria is due to rater bias.

Discussion

The previous sections of this paper have included the word "selection" in order to underline the fact that the practical need to predict to any of these criteria exists only when some sort of selection is necessary. For example, if the waiting list of an outpatient clinic is too long, selection of cases to receive treatment can be made on the basis of predicted probability of improving or terminating. If there is no need to deny treatment to anyone, knowledge of these prognostic probabilities is of no practical use. In most mental treatment centers today administrative procedures probably do not involve rejection of the patient as an alternative action, except in some outpatient clinics. Prognosis would be indispensable in the question of treatment selection, if differential effects of treatment were known; our survey has suggested that such effects have not yet been demonstrated. Thus it could be argued that prognosis is a sleeping giant at the present time, awaiting a future chance to be of service. Several other uses can be made of prognostic information, of course. Knowledge of the variables which relate to changes in duration, course, or outcome of mental illness is of theoretical importance, an aid to understanding. A second promising use has been proposed by Feldman (1952) and Zubin (1959). They recommend that in nonprognostic research prognostic status be tried as a method of classifying patients into homogeneous categories, in place of diagnosis.

Is such a suggestion tantamount to substituting a measure of severity of illness for one of type of illness? The literature survey indicates a wide variety of tests have shown positive results, with no discernible common characteristic except that they measure adequacy of functioning, directly or indirectly. The fact that the same measures do not predict for all patients may be due to differences in the type and etiology of symptoms from patient to patient; but such differences do not vitiate the possibility that when prediction occurs it is largely because the dimension of severity of illness has been accurately assessed by the test. In any case, the effect of matching groups on prognostic variables would be to control for base rate differences in improvement, a procedure which is imperative for many kinds of evaluational studies, though rarely invoked in research on therapy.

As with all predictive questions, the primary problem in prognosis is the definition of the criterion. From the point of view of decision theory, the general notion of "outcome of illness" involves assigning utility values to specific outcomes; and since cost of achieving any given outcome may be a factor, an explication of the treatment strategies is also necessary. The low interjudge reliabilities which obtain in judgments of improvement indicate that utility of outcome may differ from judge to judge. A program for achieving a more objective ranking of treatments, outcomes, or treatment-outcome combinations seems called for. Cronbach and Gleser (1957) offer a possible framework for such a program, and most of the points they make, although dealing with personnel selection, can be easily generalized to prognosis.

A frequent misinterpretation of empirical research is that it is based on no theory. In the sense of a content theory—i.e., a theory stating relationships between tangibles or concepts related to tangibles—empirical research is usually weak, though in the selection of measures some sort of rough theory has to be involved. However, empirical research often is strongly tied to a mathematical model. In prognosis the guiding model has been the linear regression model. The studies have assumed that a measurable quality exists which is linearly related to outcome. The findings in respect to performance differences between acute and chronic patients (Burdock et al., 1958) suggest that this linear model probably will have to take account of interaction effects. If so, almost all studies to date are too simple in design. They involve a one-stage decision: look at one final score per person (the final score may of course be a combination of several subsidiary scores) and assign the patient to an outcome (criterion category) by whatever rule of operation is being applied to the score. The work of Zubin's group indicates that at least a two-stage decision process is needed: (*a*) a score is obtained to decide which of several operations will be applied to a second score, and (*b*) the second score is used to assign patients to the criterion category. Indeed there is no reason why tests should not be useful as a basis for deciding what operational rule to apply to other data. The variables which appear to have the strongest relationship to outcome have been nontest variables: severity and duration of illness, acuteness of onset, degree of precipitating stress, etc. A possible direction of research might be to use tests to increase the validity of the nontest variables, either by trying to find tests which tap interactions, or which correlate with the error term in the psychiatric predictor. This latter approach has not been tried in prognosis, but it has been used with some success in personnel selection (Fulkerson, 1959; Ghiselli, 1956). A third suggested avenue of research would be to apply nonlinear or configurational models to prognostic data. The general point to be made is that prognosis research seems to require a different, more complex, mathematical model, and thus a more complex research design, than has been generally used so far. Specifically the one-stage design, where a predictor is correlated with an outcome measure, would appear to be inadequate in this field.

REFERENCES

Affleck, D. C., & Mednick, S. A. The use of the Rorschach test in the prediction of the abrupt terminator in individual psychotherapy. *J. consult. Psychol.*, 1959, **23**, 125–128.

Appel, K. E., Myers, J. M., & Scheflen, A. E. Prognosis in psychiatry. *AMA Arch. Neurol. Psychiat.*, 1953, **70**, 459–468.

Auld, F., Jr., & Eron, L. D. The use of Rorschach scores to predict whether patients will continue psychotherapy. *J. consult. Psychol.*, 1953, **17**, 104–109.

Auld, F., Jr., & Murray, E. J. Content-analysis studies of psychotherapy. *Psychol. Bull.*, 1955, **52**, 377–395.

Bailey, M. A., Warshaw, L., & Eichler, R. M. A study of factors related to length of stay in psychotherapy. *J. clin. Psychol.*, 1959, **15**, 442–444.

Barron, F. An ego-strength scale which predicts response to psychotherapy. *J. consult. Psychol.*, 1953, **17**, 327–333. (a)

Barron, F. Some test correlates of response to psychotherapy. *J. consult. Psychol.*, 1953, **17**, 235–241. (b)

Barron, F., & Leary, T. F. Changes in psychoneurotic patients with and without psychotherapy. *J. consult. Psychol.*, 1955, **19**, 239–245.

Barry, J. R. The relation of verbal reactions to adjustment level. *J. abnorm. soc. Psychol.*, 1950, **45**, 647–658.

Bayard, Jean, & Pascal, G. R. Studies of prognostic criteria in the case records of hospitalized mental patients: Affective expression. *J. consult. Psychol.*, 1954, **18**, 122–126.

Becker, W. C., & McFarland, R. L. A lobotomy prognosis scale. *J. consult. Psychol.*, 1955, **19**, 157–162.

Betz, Barbara J., & Whitehorn, J. C. The relationship of the therapist to the outcome of therapy in schizophrenia. *Psychiat. res. Rep.*, 1956, No. 5, 89–105.

Bloom, B. L. Prognostic significance of the underproductive Rorschach. *J. proj. Tech.*, 1956, **20**, 366–371.

Briggs, P. F. Prediction of rehospitalization using the MMPI. *J. clin. Psychol.*, 1958, **14**, 83–84.

Burdock, E. I., Sutton, S., & Zubin, J. Personality and psychopathology. *J. abnorm. soc. Psychol.*, 1958, **56**, 18–30.

Burdock, E. I., & Zubin, J. A rationale for the classification of experimental techniques in abnormal psychology. *J. gen. Psychol.*, 1956, **55**, 35–49.

Calden, G., Thurston, J. R., Stewart, B. M., & Vineberg, S. E. The use of the MMPI in predicting irregular discharge among tuberculosis patients. *J. clin. Psychol.*, 1955, **11**, 374–377.

Carp, A. MMPI performance and insulin shock therapy. *J. abnorm. soc. Psychol.*, 1950, **45**, 721–726.

Cartwright, D. S. Success in psychotherapy as a function of certain actuarial variables. *J. consult. Psychol.*, 1955, **19**, 357–363.

Cartwright, Rosalind D. Predicting response to client-centered therapy with the Rorschach *PR* scale. *J. counsel. Psychol.*, 1958, **5**, 11–15.

Clow, H. E. The use of a prognostic index of capacity for social adjustment in psychiatric disorders. In P. H. Hoch & J. Zubin (Eds.), *Current problems in psychiatric diagnosis.* New York: Grune & Stratton, 1953. Pp. 89–106.

Cole, Mary E., Swensen, C. H., & Pascal, G. R. Prognostic significance of precipitating stress in mental illness. *J. consult. Psychol.*, 1954, **18**, 171–175.

CONRAD, DOROTHY C. The duration of the therapeutic relationship and therapists' successive judgments of patients' mental health. *J. clin. Psychol.*, 1954, **10**, 229–233.

COWDEN, R. C., DEABLER, H. L., & FEAMSTER, J. H. The prognostic value of the Bender-Gestalt, H-T-P, TAT, and Sentence Completion Test. *J. clin. Psychol.*, 1955, **11**, 271–275.

CRANDALL, A., ZUBIN, J., METTLER, F. A., & LOGAN, N. D. The prognostic value of "mobility" during the first two years of hospitalization for mental disorder. *Psychiat. Quart.*, 1954 **28**, 185–210.

CRONBACH, L. J., & GLESER, G. *Psychological tests and personnel decisions.* Urbana: Univer. Illinois Press, 1957.

DANA, R. H. The effects of attitudes towards authority on psychotherapy. *J. clin. Psychol.*, 1954, **10**, 350–353.

DECHARMS, R., LEVY, J., & WERTHEIMER, M. A note on attempted evaluations of psychotherapy. *J. clin. Psychol.*, 1954, **10**, 233–235.

ELLSWORTH, R. B., & CLARK, L. D. Prediction of the response of chronic schizophrenics to drug therapy: A preliminary report on the relationship between palmar sweat and the behavioral effects of tranquilizing drugs. *J. clin. Psychol.*, 1957, **13**, 59–61.

ELLSWORTH, R. B., & CLAYTON, W. H. Measurement of improvement in "mental illness." *J. consult. Psychol.*, 1959, **23**, 15–20.

ENDS, E. J., & PAGE, C. W. Group psychotherapy and concomitant psychological change. *Psychol. Monogr.*, 1959, **73**(10, Whole No. 480).

ESKEY, A., FRIEDMAN, GLADYS M., & FRIEDMAN, I. Disorientation as a prognostic criterion. *J. consult. Psychol.*, 1957, **21**, 149–151.

ESKEY, A., & FRIEDMAN, I. The prognostic significance of certain behavioral variables. *J. consult. Psychol.*, 1958, **22**, 91–94.

EYSENCK, H. J. The effects of psychotherapy: An evaluation. *J. consult. Psychol.*, 1952, **16**, 319–324.

EYSENCK, H. J., GRANGER, G. W., & BRENGELMANN, J. C. *Perceptual processes and mental illness.* London: Institute of Psychiatry, 1957.

FELDMAN, DOROTHY A., PASCAL, G. R., & SWENSEN, C. H. Direction of aggression as a prognostic variable in mental illness. *J. consult. Psychol.*, 1954, **18**, 167–170.

FELDMAN, M. J. A prognostic scale for shock therapy. *Psychol. Monogr.* 1951, **65**(10, Whole No. 327).

FELDMAN, M. J. The use of the MMPI profile for prognosis and evaluation of shock therapy. *J. consult. Psychol.*, 1952, **16**, 376–382.

FELDMAN, M. J. An evaluation scale for shock therapy. *J. clin. Psychol.*, 1958, **14**, 41–45.

FILMER-BENNETT, G. Prognostic indices in the Rorschach records of hospitalized patients. *J. abnorm. soc. Pscyhol.*, 1952, **47**, 502–506.

FILMER-BENNETT, G. The Rorschach as a means of predicting treatment outcome. *J. consult. Psychol.*, 1955, **19**, 331–334.

FLEISHMAN, E. A., & HEMPEL, W. E. A factor analysis of dexterity tests. *Personnel Psychol.*, 1954, **7**, 14–32.

FLEISHMAN, E. A., & HEMPEL, W. E. Factorial analysis of complex psychomotor performance and related skills. *J. appl. Psychol.*, 1956, **40**, 96–104.

FRANK, J. E., GLEIDMAN, L. H., IMBER, S. D., NASH, E. H., & STONE, A. R. Why patients leave psychotherapy. *AMA Arch. Neurol. Psychiat.*, 1957, **79**, 283–299.

FULKERSON, S. C. Individual differences in response validity. *J. clin. Psychol.*, 1959, **15**, 169–173.

GALLAGHER, J. J. The problem of escaping clients in nondirective counseling. In W. U. Snyder, *Group report of a program of research in psychotherapy.* State College: Pennsylvania State College, Psychotherapy Research Group, 1953. Pp. 21–38. (Mimeo)

GALLAGHER, J. J. Test indicators for therapy prognosis. *J. consult. Psychol.*, 1954, **18**, 409–413.

GHISELLI, E. E. Differentiation of individuals in terms of their predictability. *J. appl. Psychol.*, 1956, **40**, 374–377.

GIBBY, R. G., STOTSKY, B. A., HILER, H. W., & MILLER, D. R. Validation of Rorschach criteria for predicting duration of therapy. *J. consult. Psychol.*, 1954, **18**, 185–191.

GIBBY, R. G., STOTSKY, B. A., MILLER, D. R., & HILER, H. W. Prediction of duration of therapy from the Rorschach test. *J. consult. Psychol.*, 1953, **17**, 348–354.

GLASS, A. J., RYAN, F. J., LUBIN, A., REDDY, C. V. R., & TUCKER, A. C. Factors influencing psychiatrists in the prediction of military effectiveness. *Walter Reed Army Inst. Res. res. Rep.*, 1956, No. WRAIR-186–56.

GLEIDMAN, L. H., STONE, A. R., FRANK, D. D., NASH, E., JR., & IMBER, S. D. Incentives for treatment related to remaining or improving in psychotherapy. *Amer. J. Psychother.*, 1957, **11**, 589–598.

GLESER, GOLDINE, HADDOCK, J., STARR, P., & ULETT, G. A. Psychiatric screening of flying personnel: Inter-rater agreement on the basis of psychiatric interviews. *USAF Sch. Aviat. Med. proj. Rep.* 1954, No. 10.

GORDON, M. H., LINDLEY, S. B., & MAY, R. B. A criterion measure of within-hospital change in psychiatric illness. *J. clin. Psychol.*, 1957, **13**, 145–147.

GRAUER, D. Prognosis in paranoid schizophrenia on the basis of the Rorschach. *J. consult. Psychol.*, 1953, **17**, 199–205.

GRAYSON, H. M., & OLINGER, L. B. Simulation of "normalcy" by psychiatric patients on the MMPI. *J. consult. Psychol.*, 1957, **21**, 73–77.

GUILFORD, J. P. The structure of intellect. *Psychol. Bull.*, 1956, **53**, 267–293.

GUILFORD, J. P. Three faces of intellect. *Amer. Psychologist*, 1959, **14**, 469–479.

GUNDLACH, R. H., & GELLER, M. The problem of early termination: Is it really the terminee? *J. consult. Psychol.*, 1958, **22**, 410.

HAMMER, E. F. The role of the H-T-P in the prognostic battery. *J. clin. Psychol.*, 1953, **9**, 371–374.

HARRIS, R. E. The prediction and measurement of drug-induced psychological change. In J. O. Cole & R. W. Gerard (Eds.), *Psychopharmacology: Problems in evaluation.* (NAS-NRC Publ. No. 583) Washington, D. C.: National Academy of Sciences–National Research Council, 1959. Pp. 514–528.

HASTINGS, D. W. Follow-up results in psychiatric illness. *Amer. J. Psychiat.*, 1958, **114**, 1057–1066.

HEMPEL, W. E., & FLEISHMAN, E. A. A factor analysis of physical proficiency and manipulative skills. *J. appl. Psychol.*, 1955, **39**, 12–16.

HILER, E. W. An analysis of patient-therapist compatibility. *J. consult. Psychol.*, 1958, **22**, 341–347. (a)

HILER, E. W. Wechsler-Bellevue intelligence as a predictor of continuation in psychotherapy. *J. clin. Psychol.*, 1958, **14**, 192–194. (b)

HILER, E. W. Initial complaints as predictors of continuation in psychotherapy. *J. clin. Psychol.*, 1959, **15**, 244–245.

HOZIER, Ann. On the breakdown of the sense of reality: A study of spatial perception in schizophrenia. *J. consult. Psychol.*, 1959, **23**, 185–194.

HUSTON, P. H., & PEPERNIK, M. C. Prognosis in schizophrenia. In L. Bellak (Ed.), *Schizophrenia: A review of the syndrome.* New York: Logos, 1958. Pp. 531–546.

HYBL, A. R., & STAGNER, R. Frustration tolerance in relation to diagnosis and therapy. *J. consult. Psychol.*, 1952, **16**, 163–170.

IMBER, S. D., FRANK, J. D., GLIEDMAN, L. H., NASH, E. H., & STONE, A. R. Suggestibility, social class and the acceptance of psychotherapy. *J. clin. Psychol.*, 1956, **12**, 341–344.

IMBER, S. D., FRANK, J. D., NASH, E. H., STONE, A. R., & GLIEDMAN, L. H. Improvement and amount of therapeutic contact: An alternative to the use of no-treatment controls in psychotherapy. *J. consult. Psychol.*, 1957, **21**, 309–315.

IMBER, S. D., NASH, E. H., & STONE, A. R. Social class and duration of psychotherapy. *J. clin. Psychol.*, 1955, **11**, 281–284.

JAHODA, MARIE. *Current concepts of positive mental health.* New York: Basic Books, 1958.

KALIS, BETTY L., & BENNETT, LILLIAN F. The assessment of communication: The relation of clinical improvement to measured changes in communicative behavior. *J. consult. Psychol.*, 1957, **21**, 10–14.

KATZ, J., & SOLOMON, REBECCA Z. The patient and his experiences in an outpatient clinic. *AMA Arch. Neurol. Psychiat.*, 1958, **80**, 86–92.

KATZ, M. M., LORR, M., & RUBINSTEIN, E. A. Remainder patient attributes and their relation to subsequent improvement in psychotherapy. *J. consult. Psychol.*, 1958, **22**, 411–413.

KAUFMAN, P. Changes in the Minnesota Multiphasic Personality Inventory as a function of psychiatric therapy. *J. consult. Psychol.*, 1950, **14**, 458–464.

KEEHN, J. D. An investigation into the value of "objective test psychoticism" in predicting response to insulin coma therapy. *J. ment. Sci.*, 1955, **101**, 871–877.

KELMAN, H. C., & PARLOFF, M. B. Interrelations among three criteria of improvement in group therapy: Comfort, effectiveness, and self-awareness. *J. abnorm. soc. Psychol.*, 1957, **54**, 281–288.

KING, H. E. *Psychomotor aspects of mental disease.* Cambridge: Harvard Univer. Press, 1954.

KIRKNER, F. J., WISHAM, W. W., & GIEDT, F. H. A report on the validity of the Rorschach Prognostic Rating Scale. *J. proj. Tech.*, 1953, **17**, 465–470.

KLOPFER, B., KIRKNER, F. J., WISHAM, W., & BAKER, G. Rorschach Prognostic Rating Scale. *J. proj. Tech.*, 1951, **15**, 425–428.

KNIGHT, R. P. Evaluation of the results of psychoanalytic therapy. *Amer. J. Psychiat.*, 1941, **98**, 434–446.

KOTKOV, B., & MEADOW, A. Rorschach criteria for continuing group psychotherapy. *Int. J. group Psychother.*, 1952, **2**, 324–333.

KOTKOV, B., & MEADOW, A. Rorschach criteria for predicting continuation in individual psychotherapy. *J. consult. Psychol.*, 1953, **17**, 16–20.

KRAMER, M., & GREENHOUSE, S. W. Determination of sample size and selection of cases. In J. O. Cole & R. W. Gerard (Eds.), *Psychopharmacology: Problems in evaluation.* (NAS-NAC Publ. No. 583) Washington, D. C.: National Academy of Sciences–National Research Council, 1959. Pp. 356–371.

KRASNER, L. Studies of the conditioning of verbal behavior. *Psychol. Bull.*, 1958, **55**, 145–170.

KUBIE, L. S. Some unsolved problems of psychoanalytic psychotherapy. In F. Fromm-Reichman & J. L. Moreno (Eds.), *Progress in psychotherapy 1956.* New York: Grune & Stratton, 1956. Pp. 87–102.

LANDIS, C., & CLAUSEN, J. Changes in sensory and motor performances induced by active psychiatric treatment. *J. Psychol.*, 1955, **40**, 275–305.

LANGFELDT, G. The prognosis in schizophrenia. *Acta psychiat. neurol. Scand., Kbh.*, 1956, Suppl. 110, 1–66.

LEARY, T., & HARVEY, J. S. A methodology for measuring personality changes in psychotherapy. *J. clin. Psychol.*, 1956, **12**, 123–132.

LEVITT, E. E. The results of psychotherapy with children: An evaluation. *J. consult. Psychol.*, 1957, **21**, 189–196.

LIBO, L. M. The projective expression of patient-therapist attraction. *J. clin. Psychol.*, 1957, **13**, 33–36.

LINDEMANN, J. H., FAIRWEATHER, G. W., STONE, G. B., & SMITH, R. S. The use of demographic characteristics in predicting length of neuropsychiatric hospital stay. *J. consult. Psychol.*, 1959, **23**, 85–89.

LORR, M., KATZ, M. M., & RUBINSTEIN, E. A. The prediction of length of stay in psychotherapy. *J. consult. Psychol.*, 1958, **22**, 321–327.

LUCE, R. D., & RAIFFA, H. *Games and decisions.* New York: Wiley, 1957.

MEEHL, P. E., & ROSEN, A. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychol. Bull.* 1955, **52**, 194–216.

MILES, H. H. W., BARRABEE, E. L., & FINESINGER, J. E. Evaluation of psychotherapy with a follow-up study of 62 cases of anxiety neurosis. *Psychosom. Med.*, 1951, **13**, 83–105.

MINDESS, H. Predicting patients' responses to psychotherapy: A preliminary study designed to investigate the validity of the Rorschach Prognostic Rating Scale. *J. proj. Tech.*, 1953, **17**, 327–334.

MORAN, L. J., FAIRWEATHER, G. W., & MORTON, R. B. Some determinants of successful and unsuccessful adaptation to hospital treatment of tuberculosis. *J. consult. Psychol.*, 1956, **20**, 125–131.

MORSE, P. W. A proposed technique for the evaluation of psychotherapy. *Amer. J. Orthopsychiat.*, 1953, **23**, 716–731.

MORTON, R. B. An experiment in brief psychotherapy. *Psychol. Monogr.*, 1955, **67**(1, Whole No. 386).

MYERS, J. K., & AULD, F. Some variables related to outcome of psychotherapy. *J. clin. Psychol.*, 1955, **11**, 51–54.

NASH, E. H., JR., FRANK, J. D., GLIEDMAN, L. H., IMBER, S. D., & STONE, A. R. Some factors related to patients remaining in group psychotherapy. *Int. J. group Psychother.*, 1957, **7**, 264–274.

ORR, W. F., ANDERSON, R. B., MARTIN, M. P., & PHILPOT, D. F. Factors influencing discharge of female patients from a state mental hospital. *Amer. J. Psychiat.*, 1955, **111**, 576–582.

PARLOFF, M. B., KELMAN, H. C., & FRANK, J. D. Comfort, effectiveness, and self-awareness as criteria of improvement in psychotherapy. *Amer. J. Psychiat.*, 1954, **111**, 343–351.

PASCAL, G. R., SWENSEN, C. H., FELDMAN, D. A., COLE, M. E., & BAYARD, J. Prognostic criteria in the case histories of hospitalized mental patients. *J. consult. Psychol.*, 1953, **17**, 163–171.

PASCAL, G. R., & ZAX, M. Psychotherapeutics: Success or failure. *J. consult. Psychol.*. 1956, **20**, 325–331.

PASCAL, G. R., & ZEAMAN, JEAN B. Measurement of some effects of electroconvulsive therapy on the individual patient. *J. abnorm. soc. Psychol.* 1951, **45**, 104–115.

PETERSON, D. R. The diagnosis of subclinical schizophrenia. *J. consult. Psychol.*, 1954, **18**, 198–200. (a)

PETERSON, D. R. Predicting hospitalization of psychiatric outpatients. *J. abnorm. soc. Psychol.*, 1954, **49**, 260–265 (b)

PHILLIPS, L. Case history data and prognosis in schizophrenia. *J. nerv. ment. Dis.*, 1953, **117**, 515–525.

PUMROY, D. K., & KOGAN, W. S. A validation of measures that predict the efficacy of shock therapy. *J. clin. Psychol.*, 1958, **14**, 46–47.

RABIN, A. I., & KING, G. F. Psychological studies. In L. Bellak (Ed.), *Schizophrenia: A review of the syndrome.* New York: Logos 1958. Pp. 216–278.

RAPAPORT, G. M. "Ideal self" instructions, MMPI profile changes, and the prediction, of clinical improvement. *J. consult. Psychol.*, 1958, 22, 459–463.

RENNIE, T. A. C. Prognosis in the psychoneuroses: Benign and malignant developments. In P. H. Hoch & J. Zubin (Eds.), *Current problems in psychiatric diagnosis.* New York: Grune & Stratton, 1953. Pp. 67–79.

REZNIKOFF, M., & TOOMEY, L. C. *Evaluation of changes associated with psychiatric treatment.* Springfield: Charles C Thomas, 1959.

RIOCH, MARGARET J., & LUBIN, A. Prognosis of social adjustment for mental hospital patients under psychotherapy. *J. consult. Psychol.*, 1959, 23, 313–318.

ROBERTS, L. K. The failure of some Rorschach indices to predict the outcome of psychotherapy. *J. consult. Psychol.*, 1954, **18**, 96–98.

ROGERS, C. R., & DYMOND, ROSALIND F. (Eds.) *Psychotherapy and personality change.* Chicago: Univer. Chicago Press, 1954.

ROGERS, L. S., & HAMMOND, K. R. Prediction of the results of therapy by means of the Rorschach test. *J. consult. Psychol.*, 1953, **17**, 8–15.

ROGERS, L. S., KNAUSS, JOANNE, & HAMMOND, K. R. Predicting continuation in therapy by means of the Rorschach test. *J. consult. Psychol.*, 1951, **15**, 368–371.

ROSEN, E. Ethnocentric attitude changes and rated improvement in hospitalized psychiatric patients. *J. clin. Psychol.*, 1954, **10**, 345–350.

ROSENBAUM, M., FRIEDLANDER, JANE, & KAPLAN, S. M. Evaluation of results of psychotherapy. *Psychosom. Med.*, 1956, **18**, 113–132.

ROSENBERG, S. The relationship of certain personality factors to prognosis in psychotherapy. *J. clin. Psychol.*, 1954, **10**, 341–345.

ROSENTHAL, D. Changes in some moral values following psychotherapy. *J. consult. Psychol.*, 1955, 19, 431–436.

ROSENZWEIG, S. A transvaluation of psychotherapy: A reply to Hans Eysenck. *J. abnorm. soc. Psychol.*, 1954, 49, 298–304.

RUBINSTEIN, E. A., & LORR, M. A comparison of terminators and remainers in outpatient psychotherapy. *J. clin. Psychol.*, 1956, **12**, 345–349.

SCHAFFER, L., & MYERS, J. K. Psychotherapy and social stratification: An empirical study of practice in a psychiatric out-patient clinic. *Psychiatry*, 1954, **17**, 83–93.

SCHOFIELD, W. Changes in response to the Minnesota Multiphasic Personality Inventory following certain therapies. *Psychol. Monogr.*, 1950, **64**(5, Whole No. 311).

SCHOFIELD, W. A further study of the effects of therapies on MMPI responses. *J. abnorm. soc. Psychol.*, 1953, **48**, 67–77.

SCHOFIELD, W., & BRIGGS, P. F. Criteria of therapeutic response in hospitalized psychiatric patients. *J. clin. Psychol.*, 1958, **14**, 227–232.

SCHOFIELD, W., HATHAWAY, S. R., HASTINGS, D. W., & BELL, DOROTHY M. Prognostic factors in schizophrenia. *J. consult. Psychol.* 1954, **18**, 155–166.

SCOTT, W. A. Research definitions of mental health and mental illness. *Psychol. Bull.*, 1958, **55**, 29–45.

SEASHORE, R. A., BUXTON, C. E., & MCCOLLUM, I. N. Multiple factorial analysis of fine motor skills. *Amer. J. Psychol.*, 1940, **53**, 251–259.

SEEMAN, J. Counselor judgments of therapeutic process and outcome. In C. R. Rogers & R. F. Dymond (Eds.), *Psychotherapy and personality change.* Chicago: Univer. Chicago Press, 1954. Pp. 99–108.

SNYDER, W. U. *Group report of a program of research in psychotherapy.* State College: Pennsylvania State College, Psychotherapy Research Group, 1953. (Mimeo)

SONDER, SYLVIA L. Perceptual tests and acute and chronic status as predictors of improvement in psychotic patients. *J. consult. Psychol.*, 1955, 19, 387–392.

STANDAL, S. W., & VAN DER VEEN, F. Length of therapy in relation to counselor estimates of personal integration and other case variables. *J. consult. Psychol.*, 1957, 21, 1–9.

STAUDT, VIRGINIA, M., & ZUBIN, J. A biometric evaluation of the somatotherapies in schizophrenia. *Psychol. Bull.*, 1957, **54**, 171–196.

STIEPER, D. R., & WIENER, D. N. The problem of interminability in outpatient psychotherapy. *J. consult. Psychol.*, 1959, **23**, 237–242.

STILSON, D. W., MASON, D. J., GYNTHER, M. D., & GERTZ, B. An evaluation of the comparability and reliabilities of two behavior rating scales for mental patients. *J. consult. Psychol.*, 1958, 22, 213–216.

STONE, BETH. Prognostic factors in schizophrenia. *J. consult. Psychol.*, 1959, **23**, 279.

STORROW, H. A. The measurement of change in psychotherapy. *Scientific Papers of the 115th Annual Meeting of the American Psychiatric Association*, 1959, 41–42. (Abstract)

STORROW, H. A. The measurement of outcome in psychotherapy. *AMA Arch. gen. Psychiat.*, 1960, 2, 142–146.

STOTSKY, B. A. A comparison of remitting and non-remitting schizophrenics on psychological tests. *J. abnorm. soc. Psychol.*, 1952, **47**, 489–496.

STOTSKY, B. A. How important is psychotherapy to the hospitalized psychiatric patient. *J. clin. Psychol.*, 1956, **12**, 32–36. (a)

STOTSKY, B. A. Vocational tests as measures of performance of schizophrenics in two rehabilitation activities. *J. clin. Psychol.*, 1956, **12**, 236–242. (b)

STRUPP, H. H. The psychotherapists' contribution to the treatment process. *Behav. Sci.*, 1958, **3**, 34–67.

SULLIVAN, P. L., MILLER, CHRISTINE, & SMELSER, W. Factors in length of stay and progress in psychotherapy. *J. consult. Psychol.*, 1958, **22**, 1–9.

SWENSON, C. H., & PASCAL, G. R. A note on the Bender-Gestalt test as a prognostic indicator in mental illness. *J. clin. Psychol.*, 1953, **9**, 398.

SWENSEN, C. H., JR., & PASCAL, G. R. Duration of illness as a prognostic indicator in mental illness. *J. consult. Psychol.*, 1954, **18**, 363–365. (a)

SWENSEN, C. H., JR., & PASCAL, G. R. Prognostic significance of type of onset of mental illness. *J. consult. Psychol.*, 1954, **18**, 127–130. (b)

TAULBEE, E. S. Relationship between certain personality variables and continuation in psychotherapy. *J. consult. Psychol.*, 1958, **22**, 83–89.

TAYLOR, D. M. Changes in the self concept without psychotherapy. *J. consult. Psychol.* 1955, **19**, 205–209.

TAYLOR, J. W. Relationship of success and length in psychotherapy. *J. consult. Psychol.*, 1956, **20**, 332.

THORNE, F. C. The prognostic index. *J. clin. Psychol.*, 1952, **8**, 42–45.

THURSTONE, L. L. *A factorial study of perception.* Chicago: Univer. Chicago Press, 1944.

ULLMAN, L. P. Selection of neuropsychiatric patients for group psychotherapy. *J. consult. Psychol.*, 1957, **21**, 277–280.

VENABLES, P. H., & TIZARD, J. Paradoxical effects in the reaction time of schizophrenics. *J. abnorm. soc. Psychol.*, 1956, **53**, 220–224. (a)

VENABLES, P. H., & TIZARD, J. Performance of functional psychotics on a repetitive task. *J. abnorm. soc. Psychol.*, 1956, **53**, 23–26. (b)

VERNIER, C. M., WHITING, J. F., & MELTZER, M. L. Differential prediction of a specific behavior from three projective techniques. *J. consult. Psychol.*, 1955, **19**, 175–182.

VINSON, D. B., JR. Response to electroshock therapy as evaluated by mirror drawing. *J. clin. exp. Psychopath.*, 1952, **13**, 201–210.

VOSBURG, R. Some remarks on psychotherapy as reflected in hospital charts. *Psychiat. Commun.*, 1958, **1**, 151–160.

WECHSLER, D. *The measurement and appraisal of adult intelligence.* Baltimore: Williams & Wilkins, 1958.

WIENER, D. N. The effect of arbitrary termination on return to psychotherapy. *J. clin. Psychol.*, 1959, **15**, 335–338.

WILLIAMS, R. J., & MACHI, V. S. An analysis of interperson correlations among thirty psychotics. *J. abnorm. soc. Psychol.*, 1957, **55**, 50–57.

WINDER, A. E., & HERSKO, M. The effect of social class on the length and type of psychotherapy in a Veterans' Administration mental hygiene clinic. *J. clin. Psychol.*, 1955, **11**, 77–79.

WINDER, C. L. Psychotherapy. *Annu. Rev. Psychol.*, 1957, **8**, 309–330.

WINDLE, C. Psychological tests in psychopathological prognosis. *Psychol. Bull.*, 1952, **49**, 451–482.

WINDLE, C. Test-retest effect on personality questionnaires. *Educ. psychol. Measmt.*, 1954, **14**, 617–633.

WINDLE, C., & HAMWI, VIOLET. An exploratory study of the prognostic value of the Complex Reaction Time Test in early and chronic psychotics. *J. clin. Psychol.*, 1953, **9**, 156–161.

WIRT, R. D. Further validation of the ego-strength scale. *J. consult. Psychol.*, 1955, **19**, 444.

WIRT, R. D. Actuarial prediction. *J. consult. Psychol.*, 1956, **20**, 123–124.

ZUBIN, J. A biometric model for psychopathology. In R. A. Patton (Ed.), *Current trends in the description and analysis of behavior.* Pittsburgh: Univer. Pittsburgh Press, 1958. Pp. 22–47.

ZUBIN, J. Role of prognostic indicators in the evaluation of therapy. In J. O. Cole & R. W. Gerard (Eds.), *Psychopharmacology: Problems in evaluation.* (NAS-NRC Publ. No. 583) Washington, D. C.: National Academy Sciences–National Research Council, 1959. Pp. 343–355.

ZUBIN, J., & WINDLE, C. Psychological prognosis of outcome in the mental disorders. *J. abnorm. soc. Psychol.*, 1954, **49**, 272–281.

ZUBIN, J., WINDLE, C., & HAMWI, V. Retrospective evaluation of psychological tests as prognostic instruments in mental disorders. *J. Pers.* 1953, **21**, 342–355.

(Received April 25, 1960)

Psychological Bulletin
1961, Vol. 58, No. 3, 205–217

COMPLEX SOUNDS AND CRITICAL BANDS[1]

BERTRAM SCHARF
Northeastern University

Studies of the responses of human observers to bands of noise and other complex sounds have led to the measure of what appears to be a basic unit of hearing, the critical band. When the frequency spectrum of a stimulating sound is narrower than the critical band, the ear reacts one way; when the spectrum is wider, it reacts another way. For example, experiments show that at values less than the critical bandwidth, both loudness and absolute threshold are independent of bandwidth; only when the critical bandwidth is exceeded do the loudness and the absolute threshold increase with the width (Gässler, 1954; Zwicker & Feldtkeller, 1955; Zwicker, Flottorp, & Stevens, 1957).

The critical band has also been measured in experiments on auditory discriminations that seem to depend upon phase (Zwicker, 1952) and in experiments on the masking of a narrow band of noise by two tones (Zwicker, 1954). In all four types of experiment—loudness, threshold, sensitivity to phase, and two-tone masking—the value of the critical band is the same function of its center frequency. The values of the critical band, as a function of the frequency at the center of the band, are given by the top curve in Figure 1. The ordinate gives the width (ΔF), in cycles per second, of the critical band; the abscissa gives the center frequency. As the frequency at the center of a complex sound increases, the critical band that is measured around the center frequency becomes wider.

Not only does the critical band have the same values when measured for several kinds of auditory response, it is also independent of such stimulus parameters as the number of components in the complex (Scharf, 1959b) and the sound pressure level (Feldtkeller, 1955; Feldtkeller & Zwicker, 1956).

Prior to the experimental measures of the critical band, Fletcher (1940) had hypothesized the existence of a critical band for masking. He suggested that when a white noise just masks a tone, only a relatively narrow band of frequencies surrounding the tone does the masking, energy outside the band contributing little or nothing. Although attempts to test this hypothesis remain inconclusive, investigators (Bilger & Hirsh, 1956; Hawkins & Stevens, 1950) have been able to calculate values for the width of these hypothetical masking bands by assuming that the masking band and the just-masked tone have the same intensity. The calculated values, which are labeled "critical ratios" in Figure 1, are smaller for the masking band than for the critical band as measured in the experiments cited above. As we shall see, this discrepancy is more apparent than real.

[1] This paper was completed while the author held a research grant, B2223, from the National Institute of Health, United States Public Health Service. The author wishes to thank S. S. Stevens for detailed criticism and advice in the final preparation of the manuscript. A. B. Warren and H. S. Zamansky are also thanked for reading and making suggestions about the final draft.

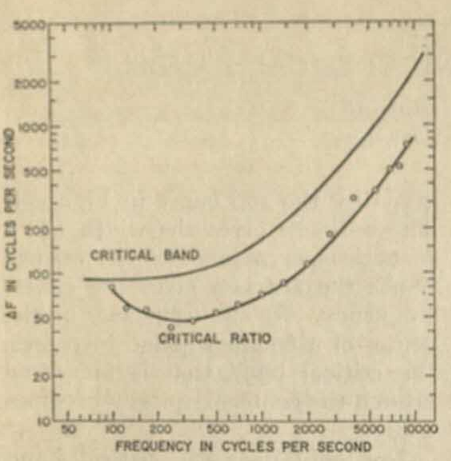


FIG. 1. The width, ΔF, of the critical band and of the critical ratio as a function of the frequency at the center of the band. (The ordinate gives the width, in cycles per second, of the critical band—and of the critical ratio—for the center frequencies shown on the abscissa. The top curve gives the values for the critical band which are based upon direct measurements in four types of experiment; the bottom curve gives the values for the critical ratio which are calculated from measurements of the masked threshold for pure tones in white noise. The points on the bottom curve are from Hawkins and Stevens—1950. This figure is adapted from an article by Zwicker, Flottorp, and Stevens—1957, p. 556—which contains also a table of critical-band values.) (Adapted with permission of the *Journal of the Acoustical Society of America*)

Experimental Measures of the Critical Band

Four types of experiment in which critical bands have been measured are reviewed: absolute threshold of complex sounds, masking of a band of noise by two tones, sensitivity to phase differences, and loudness.

Threshold of Complex Sounds

When two tones, whose frequencies are not too far apart, are presented simultaneously, a subject may report hearing a sound even though either tone by itself is below threshold. Gässler (1954) made careful measures of this phenomenon, using many tones and systematically varying the difference in frequency, ΔF, between the lowest and highest components of the complex sounds.[2] He varied the ΔF by varying the number of equally intense tones, which were spaced at intervals of 20 cps. The number of tones was increased from 1 to 40 or until ΔF was equal to 780 cps. Each time a tone was added, the threshold for the whole complex was measured by a "tracking" method (Stevens, 1958). It was necessary, of course, that all the tones in the complex have the same threshold when heard singly, for otherwise it would have been impossible to determine the precise cause of a change in the threshold for a complex whose ΔF had been increased by the addition of a tone. Thus measurements were restricted to portions of the frequency spectrum over which a subject's threshold curve was flat. In order to study other portions of the spectrum, the multitone complexes were presented against a background of white noise that had been tailored to raise the threshold for tones at all the audible frequencies to the same level, thus artificially flattening a subject's threshold curve.

Whether the background was quiet, or consisted of a noise at 0 db. SPL, at 20 db., or at 40 db., the same effect was noted: as soon as ΔF exceeded a particular value whose size depended upon the frequency at the center of the complex, the threshold for the multitone complex began to increase. Similar data were reported when bands of white noise were substituted for the multitone

[2] Two or more tones constitute a complex sound, i.e., a sound with energy at more than one frequency in contrast to a single or pure tone with most of its energy concentrated at a single frequency.

complexes. The results indicate that the *total* energy necessary for a sound to be heard remains constant so long as the energy is contained within a limiting bandwidth. Although differences between the two observers in these experiments were sometimes of the order of 40%, the average size of the limiting bandwidths for both multitone complexes and bands of noise is approximated by the critical-band curve of Figure 1.[3]

Two-Tone Masking

The masking of a narrow-band noise by two tones provided a second measure of the critical band. Using a tracking method, Zwicker (1954) measured the threshold of a narrow-band noise in the presence of two tones, one on either side of the noise. Increasing the difference in frequency, ΔF, between the two tones left the masked threshold for the noise unchanged until a critical ΔF was reached, whereupon the threshold fell sharply and, in general, continued to fall as ΔF was increased further. The two subjects who served in this experiment showed the same drop in threshold at approximately the same ΔF for a given center frequency regardless of the SPL of the masking tones. The critical-band curve of Figure 1 gives the approximate values of ΔF at which the masking effect of two tones is sharply reduced.

Sensitivity to Phase

The critical band is also relevant to phase sensitivity, measured by a comparison between the ear's ability to detect amplitude modulation (AM) and its ability to detect frequency modulation (FM). This procedure requires some explanation.

When the *amplitude* of a tone is modulated—i.e., alternately increased and decreased—a three-tone complex is produced with the original tone (the "carrier") at the center of the complex and a tone on either side (side bands). When the *frequency* of a tone is modulated over a narrow range, a three-tone complex is also produced.[4] The only important difference between the three-tone complex that is produced under AM and the complex that is produced under FM concerns the phase relations among the components. Consequently, any difference in the ear's sensitivity to AM and FM would presumably depend upon these phase relations.

Zwicker (1952) found, indeed, that in order for a subject to just hear a difference between a modulated and a pure, unmodulated tone, a smaller amount of AM is required than FM. The ear is more sensitive to AM than to FM, however, only at low rates of modulation. As the rate of modulation is increased, the difference in sensitivity to AM and FM gradually disappears. How do these results pertain to the critical band? The rate at which a tone is modulated determines the frequency separation, ΔF, between the side bands of the three-tone complex produced under the modulation. It turns out that the rate of modulation at which AM and

[3] Gässler (1954) measured a critical band of 165 cps at 1000 cps. Garner (1947) had written earlier that "the best estimate . . . is that a band of frequencies no wider than 175 cps around 1000 cps is necessary if temporal integration of acoustic energy is to be perfect" (p. 813). His estimate was based upon measurements of the threshold changes for a wide-band noise, an unfiltered 1000-cycle tone, and a filtered 1000-cycle tone as a function of bandwidth which was varied by varying the duration of the signal.

[4] For a lucid discussion of the intricacies of modulation, consult Stevens and Davis (1938, pp. 225–231).

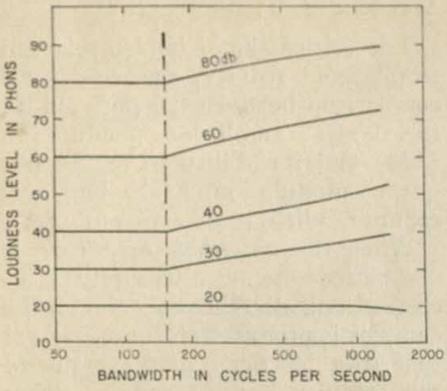


FIG. 2. The loudness level of a band of noise centered at 1000 cps measured as a function of the width of the band. (The parameter is the effective SPL of the noise. The dashed line shows that the bandwidth at which loudness begins to increase is the same at all the levels tested. This figure is adapted from the book, *Das Ohr als Nachrichtenempfänger*, by Feldtkeller and Zwicker—1956, p. 82.) (Adapted with permission of S. Hirzel Verlag)

FM become equally difficult to detect corresponds to values of ΔF that are essentially the same as the critical-band values given in Figure 1. Zwicker's investigation showed, moreover, that the critical band determined by phase sensitivity is independent of the SPL of the modulated tone and varies only as a function of the frequency of the "carrier" which lies, of course, at the center of the band.

Since the complexes produced under AM and those produced under FM differ primarily with respect to phase relations, the ear may be able to detect AM more easily than FM at low rates of modulation because it is more sensitive to the kind of phase relations that occur under AM. The ear seems to be sensitive to the phase relations, however, only when the ΔF of the complex is less than a critical band. When ΔF is greater than a critical band, there is no difference in sensitivity to AM and FM, implying that, beyond the critical band, the phase relations within the complex no longer serve as a significant cue in the detection of modulation.

Loudness of Complex Sounds

The critical band has been measured most thoroughly in studies of the loudness of complex sounds as a function of bandwidth. Zwicker and Feldtkeller (1955) demonstrated that the loudness of a white noise is independent of bandwidth until the critical band is exceeded, whereupon the loudness begins to increase. Their procedure was straightforward. They presented a band of filtered white noise and a comparison tone alternately through a single earphone. The subject adjusted the intensity of the tone until the tone and the noise sounded equally loud. The overall SPL of the noise was held constant; only the bandwidth was varied from judgment to judgment. (Zwicker and Feldtkeller did not report the number of subjects or the amount of variability; probably only a few, well-trained subjects were used and the variability was small.) Figure 2 shows what happens to the loudness of a band of noise when its width is increased. These curves are for bands centered at 1000 cps, which was the geometric mean of the two half-power points. At all the SPLs tested, from 30 to 80 db., the loudness of the noise remains constant and the curve is flat up to a bandwidth of about 160 cps, whereupon the loudness begins to increase. Within the critical band, the noises are as loud as a tone of equal intensity, having the same frequency as the center of the band. Functions similar in shape to those in Figure 2 were generated for bands centered at

500, 2000, and 4000 cps. The bandwidth at which loudness begins to increase defines the critical band for loudness, which was found to have approximately the same values as had been measured for threshold, two-tone masking, and phase sensitivity (see Figure 1).

Zwicker and Feldtkeller studied continuous spectra, i.e., noises that have energy at every frequency between the cutoff points. Bauch (1956) studied line spectra, i.e., sounds that have energy at two or more separate frequencies. He measured the loudness of three-tone complexes, produced by amplitude modulation, as a function of the difference, ΔF, in cps between the lowest and highest components of the complex. Bauch obtained the same results with three-tone complexes centered at various frequencies as Zwicker and Feldtkeller had obtained with bands of noise. For values of ΔF less than a critical band, loudness is constant except when ΔF is so small that beats are heard. The loudness begins to increase as a function of ΔF only when ΔF exceeds the critical band.

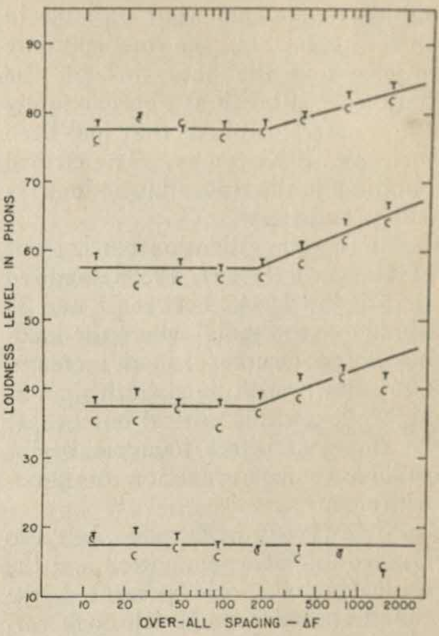


FIG. 3. The dependence of the loudness of a four-tone complex, centered at 1000 cps, on spacing and level. (Each point represents the median of two judgments by each of 10 listeners. The symbol T means the comparison tone was adjusted; C means the complex was adjusted. This figure is from Zwicker, Flottorp and Stevens—1957, p. 550.) (Reproduced with permission of the *Journal of the Acoustical Society of America*)

At the time that the critical band was being mapped out in Germany at the Technischen Hochschule Stuttgart (Bauch, 1956; Gässler, 1954; Zwicker, 1952, 1954; Zwicker & Feldtkeller, 1955) some of us at the Psycho-Acoustic Laboratory at Harvard were puzzled by our failure to find an increase in the loudness of a four-tone complex as a function of ΔF. We had assumed that loudness summation begins as soon as ΔF is increased. We were, however, studying four-tone complexes whose ΔFs were smaller than a critical band. When reports of the critical band came from Germany, our results began to make sense and, indeed, agreed well with those being obtained across the sea. The experiments were continued at Harvard by S. S. Stevens with G. Flottorp from Norway and E. Zwicker from Germany (Zwicker et al., 1957). Four-tone complexes and bands of white noise, at various center frequencies and various SPLs, were studied. In these experiments, 16 to 22 untrained subjects sometimes adjusted the complex sound and sometimes adjusted the comparison until the two were equally loud. Figure 3 shows a typical set of results, those for four-tonecomplexes centered at 1000 cps. Each point is the median of 20 loudness matches. Although the subjects were somewhat variable in their judgments, the medians are orderly and the lines through the data show a break at approximately the same value of ΔF that had been measured in Germany. The critical band made the transatlantic journey safely and invariantly.

Another investigation carried out at Harvard (Scharf, 1959a) showed that at low levels, between 5 and 35 db. above threshold, where the loudness of a complex sound increases more slowly with bandwidth than at higher levels, the critical band must be exceeded before loudness begins to change as a function of bandwidth.

Niese (1960), in Dresden, has also studied loudness summation and the critical band. He presented the sound stimuli not only through earphones (as in all the previous experiments) but also through a loudspeaker in a free field, i.e., in an anechoic room where sounds are almost completely absorbed by specially constructed walls. The results for free-field listening are similar to those for earphone listening; the loudness of a band of white noise begins to increase with bandwidth when the critical band is exceeded. Niese found, however, that the loudness did not continue to increase indefinitely with bandwidth, but increased about 8 db. and then remained constant for bandwidths greater than 1000 to 5000 cps depending upon the center frequency. It may be that the loudness did not increase further because the available energy was spread to very low and very high frequencies which contributed little to the total loudness.

In other experiments, Niese (1960) tested the assumption that loudness summation is a peripheral process occurring independently in each ear. In one procedure, a band of white noise was divided in half at its center frequency; the upper half was presented through an earphone to one ear and the lower half to the other ear. The loudness of the noise in both ears did not begin to increase with bandwidth until the *overall* width exceeded a value approximately *twice* the critical band, i.e., until the noise in each ear was wider than a single critical band. In a second procedure, two narrow bands, each 100 cycles wide, were first presented together to one ear and later separately to each ear. When presented together to a single ear, the loudness of the two bands increased with the frequency separation between them. When, on the other hand, one band was presented to each ear, the loudness did not increase with the frequency separation, no matter how great it was. The loudness did not increase because the band of noise presented to each ear was never wider than a critical band; it was always 100 cycles wide. Loudness summation thus seems to depend only upon the distribution of energy in one ear, suggesting that summation takes place not at some higher level in the auditory system where nerve impulses from the two ears join, but at the periphery, probably in the inner ear.

Still another aspect of loudness summation has been recently investigated (Scharf, 1959b). The results indicate that the loudness of a complex sound remains essentially unchanged when only the number of components in the complex is varied. The loudness of the complex increases with ΔF when ΔF is greater than a critical band, but at any given value of ΔF the loudness is approximately invariant with the number of components, provided the overall sound pressure remains invariant.

The several experiments in loudness summation, along with those on threshold, two-tone masking, and phase sensitivity provide a firm body of evidence for the critical band. There remains, however, the question of the role of the critical band in the masking of pure tones by white noise.

Masking Bands

Although the empirical measures of the critical band are quite recent, the concept of a critical band was expounded some 20 years ago by Fletcher (1940) when he hypothesized that: (*a*) a pure tone that is masked by a white noise is in effect masked only by a narrow band of frequencies surrounding the tone, and (*b*) the intensity of the part of the band that does the masking is equal to the intensity of the tone.

Fletcher (1940) presented some preliminary experimental results to support his thesis, but the projected full-scale experiment has apparently not been reported. Nonetheless the concept of a critical band has become important in theories about masking. Moreover, the acceptance of Fletcher's hypotheses permits the calculation of values for the masking band from the measurement of the masking of pure tones by white noise (Hawkins & Stevens, 1950). The calculated values for the masking band turn out to be about two-and-one-half times smaller than the empirical values for the critical band, as measured in experiments on loudness, two-tone masking, etc. This discrepancy, however, may be resolved either by a modification of Fletcher's second hypothesis, or, better, by direct measurements of the masking band. Let us turn first to the indirect measurements of the masking band and the assumptions underlying them.

Indirect Measures of the Masking Band

If both Fletcher's hypotheses about the existence of a masking band and about the equality of the intensities of the tone and noise are accepted, it is possible to calculate the size of the masking band from the masked thresholds for pure tones in white noise. Only one empirical operation is necessary. The threshold for a tone is measured in the presence of a white noise. From the intensity of the just-masked tone and the intensity of the masking noise, it is fairly simple to calculate how large a band within the noise contains the same energy as the tone. The width of this band is, *by definition*, the masking band. Its width is calculated by taking the ratio of the intensity of the tone to the intensity per cycle of the noise. (Since a white noise contains all audible frequencies at equal intensity, the intensity per cycle is uniform throughout.) For example, Hawkins and Stevens (1950) found that the ratio between the intensity of a 1000-cycle tone (at its masked threshold) and the intensity per cycle of the masking noise is 63:1 or 18 db. Since the intensity in each one-cycle band of noise is 1/63 the intensity of the masked tone, a band of frequencies 63 cps wide will have an *overall* intensity equal to that of the tone. Therefore, according to the second hypothesis, the masking band is taken to be 63 cps wide for a tone of 1000 cps. Values for the masking band that are calculated in the foregoing manner will be called "critical ratios," as suggested by S. S. Stevens (see Zwicker et al., 1957).

Hawkins and Stevens measured the masked thresholds at many frequencies from 100 to 9000 cps in the presence of white noise at levels from 20 to 90 db. They found that the ratio of the intensity of a just-masked tone to the intensity per cycle of the masking noise remains constant at all noise levels except the very lowest. In other words, the critical ratio does not change as a function of the level of the masking noise. The critical ratio is, however, different at different center frequencies, as shown in Figure 1. The results of these experiments agree with similar measurements that Fletcher and Munson (1937) had made of the critical ratio for tones masked by a uniform masking noise.

Bilger and Hirsh (1956) also calculated critical ratios from masking data obtained with bands of white noise 250 mels wide. (The mel is a unit of pitch.) The substitution of a 250-mel band, which is about five times as wide as the critical ratios measured by Hawkins and Stevens, is consistent with the assumption that the energy outside the masking band contributes nothing to the masking effect. If this, Fletcher's fundamental assumption, is true the critical ratio should be the same in both experiments. The results of the two independent experiments were, in fact, in close agreement.

In all these experiments the calculated value of the critical ratio depends upon the measured value of the masked threshold which may not be very reliable. Blackwell (1953) has shown, for example, that the value obtained for a threshold depends upon the psychophysical method employed in its measurement. The congruence of the results of the several experiments tends, however, to negate this criticism. Using the reported threshold measurements, we can modify Fletcher's second asumption so that the masking band has the same values as the critical band.

Instead of assuming, quite arbitrarily, that the intensities of the masked tone and of the masking band are equal, we can just as well assume that the intensity of the masking band is two-and-one-half times as great as that of the masked tone. Over most of the frequency range, this simple modification of Fletcher's second hypothesis yields values for the masking band that are equal to the measured values of the critical band. A simple modification succeeds because, as Figure 1 shows, except for very low frequencies, the critical band and the critical ratio are the same functions of center frequency. Since this new assumption is ad hoc and arbitrary, it will probably have little appeal. What we need is a more direct and straightforward type of evidence of the existence of the masking band.

Direct Measures of the Masking Band

The direct measurement of the masking band requires the sampling of the masked threshold for tones in the presence of bands of noise of different widths. If a masking band exists, the tone should become more difficult to detect as the bandwidth of the noise is increased up to the value of the masking band. Increasing the bandwith beyond the masking band should not raise the threshold for the tone any further. (In such experiments, energy is added to the noise as the bandwidth is increased, unlike experiments on loudness summation where a constant amount of noise energy is spread over a wider frequency range in order to increase the bandwidth.) Direct measurements of this type have been reported by Fletcher (1940), Hamilton (1957), and Schafer, Gales, Shewmaker, and Thompson (1950). Some of the recent experiments suggest that the masking band is larger than the critical ratio and may approximate the critical band as measured for other auditory phenomena.

In the first and most famous of these experiments, Fletcher (1940) measured the threshold for tones of seven different frequencies ranging from 125 to 8000 cps in the presence of bands of noise of various widths. No information about subjects, apparatus, or procedure was given. The results of this admittedly preliminary experiment provided some evidence for the masking-band hypothesis; the masked threshold tended first to increase and then to remain constant as the bandwidth of the masking noise was increased. The results seemed also to justify the assumption that, within the masking band, the intensity of the noise and the just-masked tone are equal: a band of noise, 30 cps wide, just masked a tone lying at its center frequency and having the same intensity. Precise determinations of the width of the masking band were not possible, however, because the data were highly variable and only a few bandwidths had been sampled. Of bandwidths having values in the vicinity of those for the masking band, only one, 200 cps wide, was adequately sampled. Nevertheless, relying heavily upon the assumption that the masking band and the just-masked tone are equally intense and upon the threshold measurements made in the presence of wide-band noise, Fletcher suggested values for the width of the masking band. These values, which Fletcher cautioned might be wrong by a factor of two, turned out to be approximately the same as the critical ratios calculated in 1950 by Hawkins and Stevens (see Figure 1). This similarity is not surprising, for the values recommended by Fletcher were, in effect, critical ratios. While suggestive, Fletcher's results provided neither conclusive support for his hypotheses nor a solid basis for the direct measurement of the width of the masking band.

Hamilton's (1957) more recent work provides a direct and precise measure of the masking band. Measuring the masked threshold for an 800-cycle tone in the presence of bands of noise that were centered at 800 cps and that varied in width from 19 to 1100 cps, he found that up to a bandwidth of 145 cps the masked threshold increased as the width of the masking noise increased. Beyond 145 cps the threshold remained constant, indicating that the masking band at 800 cps is 145 cps wide. The critical band measured in four other types of experiment is also about 145 cycles wide at 800 cps (see Figure 1). This coincidence of values is remarkable in view of the variability inherent in these experiments and Hamilton's apparent unfamiliarity with the other measures of the critical band.

A second important result in Hamilton's experiment shows that the difference (the signal/noise ratio) between the intensity of the 800-cycle tone at its masked threshold and the overall intensity of the masking noise is not constant, even when the width of the masking noise is less than a critical band. The signal/noise ratio decreases from about 0 db. for a band 30 cps wide to almost −4 db. for the critical width of 145 cps. (Hamilton reports similar results by Bauman, Dieter, Lieberman, and Finney, 1953.) Fletcher had also found that a band 30 cps wide just masks a tone at its center when the signal/noise ratio is 0 db., i.e., when the intensities of the tone and the noise are equal. This equality at a width of 30 cps suggested that at the critical bandwidth also, the tone and noise have the same intensity. Hamilton showed, however, that at the critical bandwidth the signal/noise ratio is not the same as at 30 cps. Accordingly, Fletcher's threshold measurements for a tone in a 30-cps-wide band of noise probably lend no support to the critical-ratio hypothesis; they are, however, consistent with critical-band values for the masking band.

Although Hamilton studied only one frequency, his results provide valuable information because they are orderly and self-consistent. Probably the use of a forced-choice procedure with well-trained subjects contributed to the preciseness of the results. In contrast, Schafer et al. (1950) report a more extensive experiment whose results are difficult to interpret. They measured the masked threshold for tones in three frequency regions as a function of the bandwidth of the surrounding noise. Instead of the usual white noise, they used bands of synthetic noise composed of tones one cycle apart. Preliminary experiments indicated no important difference between these bands of synthetic noise and bands of white noise. Twenty-five subjects served in the main experiments in which a random method of limits was used to measure the masked threshold for a tone that had been matched in pitch to the masking noise. The results suggest the presence of a masking band, but since no sharp change in the masked threshold was observed as the bandwidth was increased, the width of the masking band can be estimated only approximately. In the three frequency regions that were tested, the results suggest a masking band that is larger than that given by the critical ratio, and one that could well be as large as a critical band.

Schafer et al. (1950) interpreted their results to indicate no change in the signal/noise ratio within the masking band. Hamilton (1957), on the other hand, did find a small but consistent change in the signal/noise ratio within the masking band. Since, however, Schafer's observers were too variable to permit a precise measurement of changes in the signal/noise ratio, the small difference between the results of the two experiments is probably not significant. There is also some question about what Schafer et al. measured. Their use of a tone "matched in pitch to the masking noise" may account for some of the disparity between their results and Hamilton's.

These two experiments, by Hamilton and by Schafer, seem to be the only direct tests of the masking-band hypothesis since Fletcher's original attempt. One related experiment (Webster, Miller, Thompson, & Davenport, 1952) deserves mention. A white noise with octave gaps was used to mask tones at frequencies corresponding to those in and near the gaps. The measurements of the masked thresholds seem to suggest that Fletcher's values for the masking bands are too small.

The lack of extensive tests of the masking-band hypothesis prevents a definitive statement about the validity of the hypothesis, and even less may be said about the size of the bands. Nevertheless the net impression one obtains from the literature is that a masking band does exist and that it may well be the same width as the critical band.[5]

Other Correlates of the Critical Band

We have seen that the function relating the critical band to the frequency at the center of the band is derived from four types of experiment and that the width of the masking band may be the same as that of the critical band. Of interest, also, is the resemblance that the critical-band function bears to several other functions of frequency: the place of maximal displacement on the basilar membrane, the difference limen for frequency, and the mel scale of subjective pitch. These similarities have been noted elsewhere with respect to the critical band (Zwicker et al., 1957) and also with respect to the critical ratio (Fletcher, 1940, 1953; von Békésy & Rosenblith, 1951).

Perhaps the most interesting fact about the critical band is that it seems to correspond to a constant distance of about 1.3 millimeters along the basilar membrane. The first line in Figure 4 is a slightly idealized schematization of the frequency representation on the basilar membrane. The second line shows that 24 or 25 critical bands may be represented by equal-sized segments of the membrane. The boundaries of the critical bands are not fixed, of course, since a critical band may take shape around any frequency.

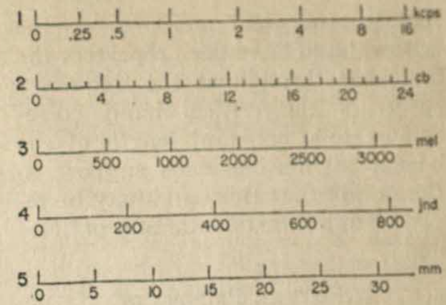


Fig. 4. Representation on the basilar membrane of (1) frequency in kilocycles, (2) critical bands, (3) pitch (Stevens & Volkmann, 1940), (4) just noticeable differences for frequency, the fifth line marks off distance in millimeters on the basilar membrane. (This figure is adapted from the book, *Das Ohr als Nachrichtenempfänger*, by Feldtkeller and Zwicker—1956, p. 60.) (Adapted with permission of S. Hirzel Verlag)

The mel and the jnd for frequency also correspond to constant distances on the basilar membrane (see the third and fourth lines in Figure 4). It is, therefore, not surprising that the critical-band function looks very much like the functions for the mel scale and the jnd scale. Measured in mels, the size of the critical band varies little, from 100 mels at low center frequencies to 180 mels at high frequencies. The mel scale is not accurate enough, however, to distinguish 100 from 180 mels at opposite ends of the scale, so that the pitch range of the critical band may, in fact, be fairly constant, perhaps approximating 150 mels.

The width of the critical band on the basilar membrane is determined from the map relating the frequency of pure tones to the position of maximal stimulation on the membrane (von Békésy, 1949). Although no di-

[5] Since the preparation of this article, Greenwood (1960) has reported an extensive study that confirms the suggestion that there is a masking band and that it is the same size as the critical band. Greenwood measured the threshold for pure tones presented in bands of white noise. He varied not only the width of the bands of noise around a given center frequency, but also the sensation level of the noise and the frequency of the masked tone. Investigating bands of noise in five regions of the spectrum, he found consistent evidence for the existence of a fairly sharp masking band approximately the same size as the critical band.

rect physiological measures of the critical band have been reported, the fact that throughout the frequency spectrum the critical band corresponds to a constant length of the basilar membrane lends support to the notion that this band may be regarded as a fundamental unit of hearing.

Future Prospects

With the experimental basis for the critical band reasonably well established, investigators are beginning to consider the relevance of the critical band to the loudness of pure tones, to temporal integration, to deafness, to speech perception, and to other auditory processes.

Zwicker (1956, 1958), for example, has argued that the loudness of an intense pure tone is a composite loudness because the displacement of the basilar membrane is spread over many critical bands. Zwicker assumes that the "loudnesses" corresponding to these critical bands summate to give the total loudness of the tone. Similar assumptions underlie Zwicker's (1958) system for the objective calculation of the loudness of a complex noise. The loudness of a noise is assumed to equal the sum of the individual loudnesses of the component critical bands after allowance for mutual masking effects among the bands.

Other investigators are studying temporal integration for short tone pulses (cf. Plomp & Bouman, 1959). Since short tone pulses are in effect multicomponent complexes whose bandwidth varies with time, the integration of energy at threshold would be expected to occur within the critical band.

Clinical use of the critical band has been attempted by deBoer (1960) in the diagnosis of hearing loss. His results suggest that the critical-band mechanism may be disturbed in certain kinds of deafness. The related problem of individual differences for the critical band has remained essentially uninvestigated except for some observations by Niese (1960) and indications from earlier data (e.g., Gässler, 1954) that the size of the critical band may vary from person to person, just as thresholds do.

Although no answers have yet come forth, phoneticists are beginning to ask about the role of the critical band in the perception of speech. Musicians may soon add their problems. The quest has begun in earnest. Now that a fundamental unit of hearing has been identified, it remains to discover its role in all the many processes called hearing.

REFERENCES

Bauch, H. Die Bedeutung der Frequenzgruppe für die Lautheit von Klängen. *Acustica*, 1956, **6**, 40–45.

Bauman, R. C., Dieter, C. L., Lieberman, W. J., & Finney, W. J. The effects of very narrow band filtering on the aural recognition of pulsed signals in noise backgrounds. *J. Acoust. Soc. Amer.*, 1953, **25**, 190. (Abstract)

Bilger, R. C., & Hirsh, I. J. Masking of tones by bands of noise. *J. Acoust. Soc. Amer.*, 1956, **28**, 623–630.

Blackwell, H. R. Psychophysical thresholds: Experimental studies of methods of measurement. *Bull. Engrg. Res. Inst., U. Mich.*, 1953, No. 36.

deBoer, E. Measurement of critical bandwidth in cases of perception deafness. *Proc. 3rd Int. Congr. Acoustics.* Amsterdam: Elsevier, 1960.

Feldtkeller, R. Ueber die Zerlegung des Schallspektrum in Frequenzgruppen durch das Gehör. *Elektrophys. Rdsch.*, 1955, **9**, 387.

Feldtkeller, R., & Zwicker, E. *Das Ohr als Nachrichtenempfänger.* Stuttgart: Hirzel, 1956.

Fletcher, H. Auditory patterns. *Rev. mod. Phys.*, 1940, **12**, 47–65.

Fletcher, H. *Speech and hearing in communication.* New York: Van Nostrand, 1953.

Fletcher, H., & Munson, W. A. Relation between loudness and masking. *J. Acoust. Soc. Amer.*, 1937, **9**, 1–10.

Garner, W. R. The effect of frequency spectrum on temporal integration of energy in the ear. *J. Acoust. Soc. Amer.*, 1947, **19**, 808–815.

Gässler, G. Ueber die Hörschwelle für Schallereignisse mit verschieden breitem Frequenzspektrum. *Acustica*, 1954, **4**, 408–414.

Greenwood, D. D. Auditory masking and the critical band. Unpublished doctoral dissertation, Harvard University, 1960.

Hamilton, P. M. Noise masked thresholds as a function of tonal duration and masking noise band width. *J. Acoust. Soc. Amer.*, 1957, **29**, 506–511.

Hawkins, J. E., & Stevens, S. S. The masking of pure tones and of speech by white noise. *J. Acoust. Soc. Amer.*, 1950, **22**, 6–13.

Niese, H. Subjektive Messung der Lautstärke von Bandpassräuschen. *Hochfrequenztech. Elektroakust.*, 1960, **69**(1), 17.

Plomp, R., & Bouman, M. A. Relation between hearing threshold and duration for tone pulses. *J. Acoust. Soc. Amer.*, 1959, **31**, 749–758.

Schafer, T. H., Gales, R. S., Shewmaker, C. A., & Thompson, P. O. Frequency selectivity of the ear as determined by masking experiments. *J. Acoust. Soc. Amer.*, 1950, **22**, 490–497.

Scharf, B. Critical bands and the loudness of complex sounds near threshold. *J. Acoust. Soc. Amer.*, 1959, **31**, 365–370. (a)

Scharf, B. Loudness of complex sounds as a function of the number of components. *J. Acoust. Soc. Amer.*, 1959, **31**, 783–785. (b)

Stevens, S. S. Problems and methods of psychophysics. *Psychol. Bull.*, 1958, **55**, 177–196.

Stevens, S. S., & Davis, H. *Hearing: Its psychology and physiology.* New York: Wiley, 1938.

Stevens, S. S., & Volkmann, J. The relation of pitch to frequency: A revised scale. *Amer. J. Psychol.*, 1940, **53**, 329–353.

von Békésy, G. The vibration of the cochlear partition in anatomical preparations and in models of the inner ear. *J. Acoust. Soc. Amer.*, 1949, **21**, 233–245.

von Békésy, G., & Rosenblith, W. A. The mechanical properties of the ear. In S. S. Stevens (Ed.), *Handbook of experimental psychology.* New York: Wiley, 1951. Pp. 1075–1115.

Webster, J. C., Miller, P. H., Thompson, P. O., & Davenport, E. W. The masking and pitch shifts of pure tones near abrupt changes in a thermal noise spectrum. *J. Acoust. Soc. Amer.*, 1952, **24**, 147–152.

Zwicker, E. Die Grenzen der Hörbarkeit der Amplitudenmodulation and der Frequenzmodulation eines Tones. *Acustica*, 1952, **2**, 125–133.

Zwicker, E. Die Verdeckung von Schmalbandgeräuschen durch Sinustöne. *Acustica*, 1954, **4**, 415–420.

Zwicker, E. Die elementaren Grundlagen zur Bestimmung der Informationskapazität des Gehörs. *Acustica*, 1956, **6**, 365–381.

Zwicker, E. Ueber psychologische und methodische Grundlagen der Lautheit. *Akust. Beih.*, 1958, **1**, 237–258.

Zwicker, E., & Feldtkeller, R. Ueber die Lautstärke von gleichförmigen Geräuschen. *Acustica*, 1955, **5**, 303–316.

Zwicker, E., Flottorp, G., & Stevens, S. S. Critical band width in loudness summation. *J. Acoust. Soc. Amer.*, 1957, **29**, 548–557.

(Received January 7, 1960)

Psychological Bulletin
1961, Vol. 58, No. 3, 218–233

PERSEVERATIVE NEURAL PROCESSES AND CONSOLIDATION OF THE MEMORY TRACE[1]

STEPHEN E. GLICKMAN
Northwestern University

For a short period between the turn of the century and the first world war, theories of perseveration figured prominently in attempts to understand many of the newly discovered phenomena of learning and forgetting. Although the exact lines of speculation varied from one writer to the next, in general, a neural fixation process was assumed to continue after the organism was no longer confronted with the stimuli to be learned. This fixation process was deemed crucial to efficient retention and interference with perseveration was presumed to have an adverse effect on an organism's ability to remember stimuli to which it had been exposed.

The first clear statement of such a consolidation theory is generally attributed to Müller and Pilzecker (1900). In order to account for the existence of retroactive inhibition, Müller and Pilzecker postulated the existence of a neural perseverative process, subject to external interference and requisite to the consolidation of the memory trace for recently acquired material. Although knowledge of the physiology of brain function was still quite limited, Müller and Pilzecker nevertheless attempted to be as precise as possible regarding the neural locus of perseveration. They rejected the notion that perseveration was in any way analogous to sense organ processes such as those believed to underlie the negative afterimage, on the grounds that these sensory processes were of too short duration. On the other hand, the perseveration which Müller and Pilzecker observed did appear to be similar to the repetitious or stereotyped behavior resulting from diseases of the subcortical motor centers. It was with these latter structures that Müller and Pilzecker associated perseverative activity.

Numerous other psychologists were concerned with perseveration theory during the early 1900s. Among these, DeCamp (1915) advanced what was probably the most detailed piece of pseudoneurological speculation:

> From the neurological standpoint, in the learning of a series of syllables, we may assume that a certain group of synapses, nerve cells, nerve paths, centres, etc., are involved. Immediately after the learning process the after-discharge continues for a short time, tending to set associations between just learned syllables. Any mental activity engaged in during this after-discharge, involving or partially involving the same neurological group, tends, more or less, to block the after-discharge, and give rise to retroactive inhibition (p. 68).

Some years previous, Sherrington (1906) had described the phenomenon of afterdischarge in spinal reflexes and discussed the blockage of such discharges by subsequent stimuli. It is interesting to note that

[1] The preparation of this paper was made possible by a postdoctoral fellowship held by the author at the Department of Physiology and Biophysics, University of Washington Medical School, during the summer of 1959, and supported by Grant 2-B5082 from the National Institute of Neurological Diseases and Blindness. The author is indebted to C. P. Duncan, S. M. Feldman, T. Kennedy, and D. Kimura for critically reading the manuscript.

this provided the theoretical model for DeCamp's view of perseverative processes in much the same manner as Sherringtonian physiology generally shaped the psychologists' conception of neural activity (see Hebb, 1951).

As a behavioral theory of retroactive inhibition, however, perseveration theory met with many difficulties, and was eventually replaced by the current concepts of associative interference (McGeoch & Irion, 1952; Osgood, 1953), although it continued to receive some limited support as a possible factor in forgetting (Woodworth, 1938). Ultimately, a perseveration theory, erected on the basis of inferences from behavior, was no longer viable once the behavioral observations were either shown to be false or explained more parsimoniously by other hypotheses. The rejuvenation of this theory awaited direct support from neurology.

Lashley (1918) once made the following comment on perseveration theory:

> If there is a gradual strengthening of associations during periods of nonpractice, there is implied a continuation of chemical changes within the nerve cells, initiated by the passage of a neural impulse through new channels and persisting for hours or even days without the influence of continued impulses. The experimental evidence upon which the belief in a gradual fixation of associations is based is far from convincing . . . it all can be explained equally well by other hypotheses and, in view of the extreme importance of the point for physiological explanation, we should be careful not to accept the assumption of a gradual setting of new functional connections until some real evidence is advanced to support it (pp. 363–364).

This healthy skepticism was certainly justified, although even at the time some physiological evidence was available to buttress perseveration theory.

Retrograde Amnesia

Shortly after the publication of Müller and Pilzecker's work, McDougall (1901) called attention to the applicability of their perseveration theory to the explanation of retrograde amnesia (RA) resulting from cerebral trauma. However, Burnham (1904) was apparently the first individual to extensively discuss the relationship between RA and perseverative "consolidation" amnesia. Burnham's paper involved an analysis of two cases of retrograde amnesia. Both of these subjects had sustained head injuries as the result of accidents and in both cases there was a loss of memory for events occurring during the period preceding the accident. As the result of his studies of these cases and of others cited by Ribot (1892), Burnham suggested that

> The fixing of an impression depends upon a physiological process. It takes time for an impression to become so fixed that it can be reproduced after a long interval; for it to become part of the permanent store of memory considerable time may be necessary. This we may suppose is not merely a process of making a permanent impression upon the nerve cells, but also a process of association, of organization of the new impressions with the old ones (p. 392).

He further speculated that: (*a*) the time required for this fixation process may vary with individuals and conditions; (*b*) shock produces its effects by arresting the fixation process in the nervous tissue; (*c*) such shock may be produced by great fatigue, excitement, unconsciousness, or narcotics; (*d*) RA is not all-or-none and the extent of the amnesia is relative to the amount of time elapsing before the fixation process is interrupted and finally (*e*) that automatic activity is an important factor in fixing impressions although it may not necessarily be directly observable in terms of movements.

These remarkable observations would appear to have been borne out by recent experiments in nearly every case and we can now advance these propositions with much more confidence.

During the first 4 decades of this century, the phenomenon of RA constituted the only direct physiological evidence for the existence of a neural fixation process. Early references to it are to be found in Ballard (1913), Pillsbury (1913), DeCamp (1915), and others. Although a complete review of this literature is beyond the scope of the present paper, it is perhaps worthwhile to examine the results of a comprehensive study by Russell and Nathan (1946). In a survey of 1,029 cases of head injury, only 133 were found to have experienced no RA whatsoever. Seven hundred and seven reported amnesia for events occurring from several seconds to 30 minutes preceding the injury, while 133 reported RA of more than 30-minutes duration. Records were unavailable with 56 patients in the sample. Russell and Nathan noted that the duration of RA is "in most cases a few moments only." Since the use of barbiturate hypnosis reduced the period of RA in only 6 of 40 cases, and produced no data suggestive of hysterical repressions, the authors conclude that loss of the material is due to a blocked perseveration process:

> It seems that the mere existence of the brain as a functioning organ must strengthen the roots of distant memories. The normal activity of the brain must steadily strengthen distant memories so that with the passage of time these become less vulnerable to the effects of head injury (p. 299).[2]

Experimentally induced RA has produced the best evidence for the existence of a consolidation process since the results would be predictable from perseveration theory, while the primary competing theory, the associative interference theory, has no explanation to offer. We will therefore turn now to a review of the various experimental procedures used to induce RA and the results obtained.

Electroconvulsive Shock

The introduction of electroshock therapy in 1937 provided both the impetus and the technical apparatus for the laboratory study of RA. Immediately after its introduction many practitioners observed that electroconvulsive shock (ECS) produced a temporary postshock amnesia which eventually shortened to a genuine RA for events immediately preceding the shock treatment. Zubin and Barrera (1941) were the first investigators to subject these observations to systematic study. They trained 10 patients in a series of paired associate lists to a criterion of two consecutive correct repetitions. Learning occurred either in the morning or evening, while the retention tests were given during the subsequent afternoon. The same subjects were used in control and experimental conditions, i.e., (*a*) with no shock intervening between learning and the retention test, and (*b*) with an ECS interpolated after the morning learning session. With no intervening shock there were significant savings between

[2] Coons and Miller (1960) have recently called attention to the possibility of sampling artifacts confounding the consolidation interpretation of clinical observations of retrograde amnesia. Thus, they have pointed out that, if an injury produces a general decrement in memory, positive evidence for memory is more likely to be secured while examining the larger time samples involved in remote memories as compared to recent memories.

learning and relearning, with an interpolated ECS there were no significant savings. A comparison between the effects of ECS on material learned the evening prior to shock with material learned the morning preceding shock indicated that recent material was more severely affected by ECS than remote material. The latter conclusion was based on rather small differences in savings scores and insufficient data are presented to permit adequate statistical evaluation. However, Flescher (1941), Williams (1950), and Cronholm and Molander (1958) have subsequently confirmed the substance of Zubin and Barrera's assertions. The various investigators using human subjects, although successfully employing ECS to interfere with memory, had not attempted to adequately define the time relations of such interference. This critically important step was taken by Duncan (1949). Duncan's procedure involved training rats to avoid shock to the feet in a shuttlebox situation. A light, turned on 10 seconds prior to grid shock, served as the conditioned stimulus (CS). The animals received one trial per day for 18 days and records were kept of the number of successful avoidance responses. Nine groups of animals were used in the study. Rats in eight of these groups received an ECS after each day's trial, the trial-ECS interval ranging from 20 seconds to 14 hours. In the remaining group, the ear clips used for delivering the ECS were applied following each day's trial but no current was passed. The results clearly indicated a deleterious effect of ECS on performance, the magnitude of the effect decreasing as the trial-ECS interval increased to produce a negatively accelerated curve. This general finding has since been confirmed by Ransmeier (1953), Thompson and Dean (1955), and Leukel (1957). All of the findings are compatible with the view that a single ECS can produce deficits in retention if delivered within 15 to 60 minutes following a learning trial. Moreover, ECS induced immediately following the learning trial effectively obliterates nearly all retention of the "learned" response. The studies following Duncan's have employed different learning tasks. Leukel (1957) and Ransmeier (1953) used maze learning situations with the ECSs being delivered at varying posttrial intervals. Thompson and his collaborators have employed a visual discrimination learning task, with avoidance of grid shock as the motivating agent. In these latter studies (Thompson, 1957a; Thompson & Dean, 1955; Thompson & Pennington, 1957) a single ECS was administered at various intervals following a series of massed trials in the apparatus. As a result of these extensive experiments, it has been determined that ECS produces greater deficits in young than adult rats (Thompson, 1958a; Thompson, Haravey, Pennington, Smith, Gannon, & Stockwell, 1958). Further, rats suffering from anoxia induced brain damage (Pennington, 1958), show greater deficits resulting from a single ECS than intact control animals. Both the findings with respect to age and those relating to brain damage are compatible with Thompson and his co-workers' (1958) hypothesis that the extent of the deficit will be proportional to the number of cortical neurons available. Pennington (1958) has alternately suggested that the results obtained with brain damaged rats may be a function of a prolonged perseveration process in these animals.

Thompson and Pennington (1957)

have also found that the memory decrement produced by a single ECS was less after spaced trials than after massed trials. This result was expected from the point of view of a perseveration theory as a joint function of "firmer fixation of the memory trace owing to a longer duration of perseveration" and "the lessened intensity of perseveration at the end of training due to dissipation of perseverative activity."

Although the empirical result of interference with performance by postlearning ECS has not been questioned, the interpretation of the results is not quite as clear. The points to be discussed below actually raise questions of interpretation which apply not only to the ECS procedures but to other interpolated physiological procedures as well.

1. The most serious alternative to a consolidation interpretation of the ECS results has been offered by Miller and Coons (1955). These investigators trained rats to eat in a runway and then shocked them while eating there. Avoidance was measured by an increased latency of approach to the eating place. ECSs were delivered to the animals at varying intervals after shock to the mouth. Miller and Coons reasoned that any aversive qualities of the ECSs might be expected to produce increased avoidance. On the other hand, if the ECS really interrupted consolidation, the subjects would show the opposite behavior, namely, approaching the food without hesitation. In this experiment no evidence was found for an attenuation of the avoidance response by the ECS, leading the authors to argue that the retardation in learning observed by Duncan (1949) was simply a function of placing the rat in a conflict situation. In a more recent set of experiments, Coons and Miller (1960) have succeeded in opposing the conflict and consolidation interpretations in a double grid-box situation similar to that used by Duncan. Here again, their results indicate that ECS may not eliminate memory but merely induce anxiety or conflict which inhibits performance of the response in question. They further buttress their contentions regarding the fear inducing qualities of ECS with observations on increased defecation, urination, and weight loss in those animals for whom the performance of an otherwise rewarded response is followed by an ECS. In both of their studies, the ECS apparently summated with the grid shock to produce a result which significantly favored a conflict as opposed to a perseveration interpretation. Observations of Gallinek (1956) suggest that analogous anxiety builds up in human beings during the course of electroshock therapy. Such an interpretation is logically possible for both the avoidance situations used by Duncan and by Thompson and Dean, and the maze learning situations used by Leukel and by Ransmeier and Gerard. The standard control for this has been to employ groups receiving painful but nonconvulsive shocks. In these cases (Duncan, 1949; Leukel, 1957; Ransmeier & Gerard, 1954) it has been found that (*a*) the decrements produced by the painful but nonconvulsive shocks are not nearly as severe as those produced by ECS at comparable intervals, and (*b*) the posttrial interval during which painful shocks produced their effect was always much shorter than that during which significant decrements could be produced by ECS. These latter control results would seem to

indicate that the ECS results are due to more than just conflict. It might be argued, however, that the ECSs are sufficiently more painful or unpleasant than the leg or tail shocks to account for the greater deficits produced by the former. Reference to human subjects would suggest that this is not the case. Patients do not necessarily report pain as an accompaniment of a properly delivered ECS (Stainbrook, 1948). In view of this, and considering that there is no consistent experimental evidence for punishment obliterating verbal material (Rapaport, 1942), it seems unlikely that the deficits observed in humans following ECS (or any cerebral trauma) can be explained purely in conflict terms. Finally, in regard to the animal literature, it seems reasonable to point out that Miller and Coons delivered a series of ECSs on successive days, whereas Thompson and his collaborators eliminated a persistently rewarded response with a single ECS.

In order to explain Thompson's results in conflict terms, one would have to assume a delay of reinforcement gradient lasting at least 60 minutes, and the build up of a significant amount of fear following a single ECS. Such assumptions, although possible, would not be easy to support at the present time. Clearly, however, other workers should carry out experiments utilizing designs similar to those employed by Miller and Coons, i.e., opposing the consolidation and conflict interpretations. The writer has used such a procedure in an experiment involving direct stimulation of the brain (Glickman, 1958) and this could easily be adapted for ECS. Moreover, the one-trial learning situation employed in this latter experiment would permit the use of a single ECS and enable an exceedingly accurate estimate of the trial-ECS interval.[3]

2. An alternate interpretation of the ECS results is also possible in those studies employing food reward. As Kohn (1951) and Berkun, Kessen, and Miller (1952) have shown, the rewarding properties of food are derived in part from stimulation of receptors within the mouth, and in part from actions within the stomach. ECS delivered shortly after a learning trial might act to prevent the perception of the feedback from the stomach and thereby cut down on the reinforcing properties of the food. In view of the relatively minor contribution of these stomach receptors, particularly in the early stages of learning, such effects are probably insignificant in the studies of Ransmeier (1953) and Leukel (1957).

3. A question has arisen about distinguishing between ECS effects on a time-limited consolidation process and the more generalized memory deficits which have been observed to follow a series of ECSs (see Stainbrook, 1946, for review). In particular, Worchel and Gentry (1950) have suggested that Duncan's (1949) finding of a limited period following learning when an ECS will be effective is a result of his failure to used massed ECSs. On the basis of some T maze data of their own, Worchel and Gentry argue that Duncan might

[3] Since this article went to press, there have been two reports of experiments in which the conflict and consolidation interpretations of "forgetting" have been opposed in one-trial learning situations. In both of these cases, in which the introduction of various chemical agents served as the interpolated procedure, the results favored a consolidation interpretation of the effects (Essman & Jarvik, 1960; Pearlman, Sharpless, & Jarvik, 1961).

have considerably extended the duration of time during which ECS would produce a deficit if if he had given a series of ECSs. Worchel and Gentry's results do not contradict the general finding that it is easier to disrupt learning in the period immediately following exposure to the learning situation. However, at the present time, the ECS data are compatible with the notion that the strengthening of memory traces is a continuous one throughout the life of the organism. For example, Brady (1951, 1952) has found evidence of spontaneous growth in the strength of a conditioned emotional response during a period of 90 days. On the basis of current evidence, one might expect that the interval following a learning trial, during which time interference with retention can be produced, is a direct function of the degree of physiological severity of the interpolated procedure.

Ultimately, there is probably some practical limit on the time interval between learning and ECS during which selective effects on retention can be produced. In addition, the effects of a series of ECSs delivered many hours or days after learning are often apparently temporary (Stainbrook, 1946). Brady (1951) reported that a series of ECSs suppressed a conditioned emotional response (CER) for a period of a month, although the habit reappeared spontaneously at the end of that time. It has also been found that the effects of a series of ECSs may be selective for emotional responses (Geller, Sidman, & Brady, 1955). On the basis of the accumulated data, it seems reasonable to suggest that ECS may affect performance in a number of ways including: (*a*) a temporary suppressor action involving those cerebral structures mediating pain or anxiety responses (such a mode of action would explain the proactive effects noted by Poschel, 1957, and Carson, 1957, on avoidance conditioning) and (*b*) a direct action on the neural circuits involved in memory which, if the learning-ECS interval is brief enough and the treatment sufficiently severe, may permanently erase the effects of such learning.

Anoxia

Hayes (1953) demonstrated equivalent retroactive effects of anoxia and ECS on maze learning in rats. He used a distributed practice procedure and administered the experimental treatment one hour after each trial. The experimental rats showed similar retardation in learning when their acquisition curves were compared with normal control animals. Hayes reports that histological examination of the brains produced no clear evidence of brain damage for any of the animals. Ransmeier and Gerard (1954) have also reported disturbances in maze learning resulting from anoxia, the magnitude of the disturbance decreasing "along characteristic curves with increasing intervals between training and experimental procedures."

Using a discrimination learning procedure, Thompson and Pryer (1956) showed that anoxia, produced by placing rats in a decompression chamber during the postlearning period, could lead to decrements in retention analogous to those produced by ECS. In a later study, Thompson (1957a) found that a 10-minute exposure to a simulated 30,000-foot altitude produced deficits equivalent to those resulting from ECS, although exposure to a 20,000-foot altitude did not produce such severe effects. Finally, Thompson

(1957a) has also reported that when an ECS was given 30 seconds posttraining, a subsequent 10-minute exposure to a simulated 30,000-foot altitude did not produce an additional deficit.

Temperature

A number of investigators have studied the effects of postlearning temperature on retention. In most of the earlier work (French, 1942; Hunter, 1932; Jones, 1943) the aim was to reduce the activity of the experimental group and thereby reduce retroactive inhibition. Considered in the light of the ECS literature, these studies are not immediately relevant to the present review because of the prolonged interval between the learning trials and the achievement of the desired temperature change.

In the most recent studies of Cerf and Otis (1957) and Ransmeier and Gerard (1954) it appears that temperature may have some effect on processes related to consolidation. The former investigators gave goldfish 10 massed trials in an avoidance situation using a shifting light as the CS. At varying intervals following the trials, (0 minute, 15 minutes, 60 minutes, or 4 hours) the body temperatures of different groups of 15 to 19 subjects were raised briefly to a point sufficient to induce heat narcosis (36.5°–37.0° C). In retention tests carried out the next day, the criterion of five consecutive correct responses in 10 trials was met by only 10.5% of the group narcotized immediately after learning, while 56.2% of the subjects paralyzed 4 hours following learning met the same criterion. The remaining two groups occupied intermediate positions. Fifty percent of a group of untreated control subjects also met the above criterion. Thus, the temperature induced narcosis produced much the same effect in the goldfish that ECS and anoxia have been found to produce in rodents. Ransmeier and Gerard (1954) did not find any evidence of retroactive effects of lowered body temperatures on retention of a maze habit in the hamster. Gerard (1955) has reported, however, that lowering the body temperature will apparently prolong the period during which an ECS may produce severe deficits. Thus, "hamsters kept cool between learning and electroshock show as great a disruption of learning at an interval of one hour as warm ones do at an interval of fifteen minutes." Evidently, temperatures sufficient to impair spontaneous activity in the brain as indicated by the EEG will not act directly to block consolidation, although they may slow down the chemical processes involved in the fixation of the trace.

Fay (1940) has reported RA in human subjects for events occurring while the patients were refrigerated, i.e., when the body temperature fell below 33.3° C. Under these circumstances, the subjects could respond to questions and carry on a conversation, although interrogation after the refrigeration procedure showed a loss of memory for the entire interchange. Such deficits could be explained in terms of an impairment of activity in those structures responsible for the consolidation process. However, alternative explanations are also possible.

Anesthesia

Leukel (1957) has reported that sodium pentothal injected intraperitoneally (IP) after each learning trial impaired acquisition in a maze in experimental rats when their time or error scores were compared with any of three control groups. Subjects in the three control groups received either: an IP injection of water following each trial, an IP injection of pentothal 30 minutes following each trial, or no injection. The scores did not differ among these latter groups. Leukel interpreted his results in terms of interruption of consolidation of the memory trace in those subjects receiving pentothal one minute after each trial.

On the other hand, Russell and Hunter (1937) and Ransmeier and Gerard (1954) have not found deficits in retention to result from postlearning barbiturate anesthesia. There are numerous differences in procedure, however, which might account for this discrepancy. For example, Russell and Hunter (1937) administered sodium amytal subcutaneously after giving their experimental subjects five massed trials in a maze. They observed no effects of the injection on subsequent retention of the maze. However, the subcutaneous route of injection undoubtedly prolonged the time before the drug took effect (in comparison with the IP route used by Leukel). In addition, the massed trials procedure used by Russell and Hunter resulted in a longer interval between learning and anesthesia than the Leukel procedure of injecting one minute following each trial.

Ransmeier and Gerard (1954) and D. Kimura and S.E. Glickman (unpublished) failed to find retention deficits as the result of anesthetizing hamsters or rats with ether following maze learning trials, or electric shock in an avoidance learning situation, respectively. These results suggest that the apparent effectiveness of barbiturates, as opposed to ether, in blocking consolidation may be due to secondary effects of the former on blood chemistry or blood pressure rather than direct synaptic interference. Barbiturate anesthetics produce many more severe blood changes than ether including reductions in blood pressure and blood sugar level (Kohn, 1950).

If anesthetics can be shown to exert reliable retroactive effects on learning, they may eventually prove useful in the localization of the neural structures crucial to consolidation. Techniques have recently been developed which permit the delivery of small quantities of various drugs to restricted sites within the brain of a "behaving" animal (Fisher, 1956; Olds & Olds, 1958). Utilizing such techniques, it should be possible to selectively and temporarily block activity in various cerebral structures during the period immediately following exposure to the learning situation and thereby determine which structures, if any, are crucial to the consolidation process.

Brain Stimulation

Mahut (1958), Glickman (1958), and Thompson (1958b) have reported retroactive effects of brain stimulation on learning. The stimulation was accomplished with chronically implanted electrodes which permit the animal freedom of movement in the learning situation, but enable the experimenter to deliver a small electric current to particular sites within the CNS at any chosen time. This technique enables much more specific delimitation of the structures involved in the presumed fixation process than, for example, ECS or anoxia. However, in the studies carried out thus far, there are numerous factors which serve to complicate comparisons among the studies, as well as to rule out any simple "consolidation" interpretation of the results.

Mahut (1958) tested the effects of stimulation of the nonspecific thalamic nuclei on the performance of rats in a Hebb-Williams maze. Brief bursts of 60-cycle, sine wave, 0.25-volt stimulation were delivered through implanted electrodes while the rat was eating in the goal box. Such stimulation produced poorer performance in the maze, when the error scores of these "thalamic" rats are compared with those of rats receiving either no stimulation or similar stimulation of the midbrain tegmentum. The possibility exists in this study that the effects of stimulation were not retroactive but contemporary, i.e., interfered with the animals' registration of the food reward. This might be clarified by a parametric investigation of the time interval between learning trial and stimulation, following the design of the ECS studies (Duncan, 1949; Thompson & Dean, 1955).

Glickman (1958) examined the effects of stimulation of the midbrain portion of the arousal system on the acquisition of an avoidance habit in the rat. Three 20-second bursts of stimulation, at considerably higher voltages than those used by Mahut (1958), were delivered immediately following shock to the mouth while the subjects were eating at a distinctive metal food spout. In retention tests carried out the following day, the animals who had received reticular stimulation after mouth-shock showed less avoidance of the spout (more eating behavior) than control animals not receiving brain stimulation. The interpretation of this study is also complicated due to the particular characteristics of the Hudson (1950) one-trial learning apparatus which evidently lead to a portion of the avoidance response being learned in the postshock period. Hudson has reported that the visual scanning which the animal engages in during the postshock period will reinforce the avoidance response. Thus, it is conceivable that the reticular stimulation could have simply interfered with an ongoing visual process rather than retroactively interfering with previous learning.

Thompson (1958b), in an ingeniously designed study which permits him to use each animal in a variety of experimental conditions, has reported interference with the performance of cats in an alternation task as the result of intracranial stimulation. This effect was achieved with bilateral stimulation of the caudate nucleus following each trial in a modified Wisconsin General Test Apparatus. Similar stimulation of the midbrain tegmentum did not produce the retroactive effect, although it did interfere with performance when the stimulation was delivered either before or during a given trial. In this case, the interpretation of the retroactive disruptive effects of caudate stimulation is complicated by the possible reinforcing properties of this stimulation. Brady, Boren, Conrad, and Sidman (1957) have reported positively reinforcing consequences of caudate stimulation in the cat. It seems plausible that stimulation in this region, following a particular response, would favor repetition of that response and might act in opposition to any alternation habit. Such an explanation might be an alternative to postulating interference with a perseveratory process. Since it is possible to check on the rewarding properties of electrical stimulation, using a self-stimulation situation such as that used by Olds and Milner (1954), this factor could be easily controlled in future studies. In regard to the lack of effect of tegmental stimulation, this may be explicable in terms of the extensive functional localization of reinforcement pathways which appears to exist in that region (Glickman, 1960; Olds & Peretz, 1959). Olds[4] has suggested that the interference produced by intracranial stimulation in learning situations may be directly related to the reinforcing qualities of the stimulation.

There are numerous studies demonstrating interference with learning as a result of intracranial stimulation (see Zeigler, 1957, for review). However, most of these are not directly interpretable in terms of retroactive interference because the stimulation is delivered during the actual performance of the task. Nevertheless, as Thompson (1958b) suggests, interference with consolidation may be at least a partial explanation of the deficits observed by Rosvold and Delgado (1956) coincident with caudate stimulation. Similarly, Burns and Mogenson (1958) and Burns and Stackhouse (1959) have reported deficits in the acquisition of a bar pressing habit in the Skinner Box resulting from a cortical stimulation. As Burns and Stackhouse note, these results are compatible with a perseveration hypothesis.

Physiological Substrate of Consolidation

Stellar (1957) has pointed out that physiological data have recently accumulated which tend to support the existence of a system within the brain responsible for the permanent fixation of memory traces. Milner and Penfield (1955) and Scoville and Milner (1957) have reported cases of temporal lobe ablation in man which produced severe impairment of the ability to acquire new material postoperatively, although preoperatively acquired material was retained. Although the crucial structures have not yet been definitely localized, the hippocampus and amygdala appear to be directly involved. Along similar lines Brady, Schreiner, Geller, and Kling (1954) found interfering effects of amygdalectomy on the acquisition of an avoidance response in cats, although the same lesions produced in cats which had already acquired the habit led to no disturbance in performance. The anatomical and physiological data suggest numerous pathways through which these relatively primitive temporal lobe structures could exert widespread effects on the remainder of the brain (Adey, Merrillees, & Sunderland, 1956; Green & Adey, 1956). For example, the continued action of these temporal lobe regions may be necessary to the proper regulation of firing in the nonspecific arousal system, which in turn apparently exerts considerable influence on cortical activity (Magoun, 1958).

The existence of structures within the brain which are crucial to the fixation of memory traces is not restricted to the vertebrate orders. Boycott and Young (1950) have identified a cerebral structure (the vertical lobe) requisite for fixation of visual memory in the octopus, and apparently homologous in function to the temporal lobe structures found in the higher vertebrates. Thus, removal of the vertical lobe drastically impairs the ability of the animal either to acquire a new visual discrimination habit (motivated by a combination of food and electric shock), or to retain such a habit for any length of time following training. The nervous system of the octopus differs widely from the vertebrate

[4] J. Olds, personal communication, 1959.

nervous system. However, the appearance of a specialized fixation mechanism in both invertebrates and vertebrates suggests that there is some evolutionary utility in a dual process underlying memory function.

At a more molecular level, the most widespread hypothesis concerning the substrate of consolidation predicates its dependence on reverberatory circuits. This idea has its origins in the anatomical demonstrations of Lorente de No (1938) and has been subscribed to in varying forms by Hebb (1949), Young (1953), and Gerard (1955). The basic supposition is that reverberatory activity maintains the memory until the permanent changes underlying fixation of the trace have been completed. This dual process hypothesis of memory fixation has the advantage of explaining why interference with neural activity immediately after "learning" blocks retention while similar procedures instituted at a later time do not. One group of studies which may be directly relevant to the reverberatory circuit hypothesis of consolidation has been carried out by B. D. Burns and his co-workers (Burns, 1954, 1958). Burns has developed a technique which allows the isolation of small areas of cortex from the remainder of the brain, while leaving the blood supply to the area relatively unaffected. He has extensively studied the electrical activity of these isolated slabs in response to direct electrical stimulation. Interestingly enough, he has found: that a single train of pulses can initiate bursts of activity in one of these preparations lasting for 30 minutes or more; that such bursts of activity can be blocked by a subsequently applied electrical stimulus; that such activity becomes easier to evoke with repeated applications of the stimulus; and that the burst activity is apparently due, in part to reverberatory activity among groups of neurons, and in part to differential rates of depolarization within various segments of individual neurons. These first three observations certainly coincide with what one would expect if such a process underlay consolidation. However, it is necessary to be cautious in generalizing from the type of activity observed in these special preparations to that occurring in the intact brain. Burns (1958) himself has rejected these preparations as a general model for memory on the grounds that such circuits would be too susceptible to external interference. However, this is one aspect of the data which makes Burns' findings so attractive as a model of the first phase of a dual process theory of memory, susceptibility of learned material to interference providing the main behavioral evidence for the existence of a consolidation process.

Finally, moving to a still more molecular analysis of the problem, it is reasonable to inquire about the specific changes which might be produced by some sort of perseverative process. Nearly all investigators have at this level proposed some sort of growth process or chemical change at the synapse. In this respect, our ideas have changed little from those of 1929 when Lashley wrote:

> We have today an almost universal acceptance of the theory that learning consists of modification of the resistance of specific synapses within definite conduction units of the nervous system.

After expressing numerous reservations about the adequacy of this assumption, Lashley concluded by noting that:

> The synapse is, physiologically, a convention to describe the polarity of conduction in the

> nervous system of higher animals, together with some similarities of function in the central nervous system and neuromuscular junction. That these functions are due to the action of the intercellular membranes has not been directly demonstrated (p. 127).

Here again, recent neurophysiological progress tempers Lashley's skepticism. The synapse is no longer a "convention" but a point-at-able structure which can be photographed and studied with the electron microscope (Palay, 1956). Further, as Lloyd (1949) and Eccles (1953) have shown, the rapid firing of impulses across synaptic junctions can result in increased excitability of these synapses for periods lasting from minutes to hours. There is general agreement that this increased excitability results from the firing of presynaptic fibers, although it is not yet clear whether this is in turn due to an actual change in the dimensions of the synaptic knobs as suggested by Eccles (1953, 1957) or if an alternate explanation, e.g., Lloyd (1949), may suffice. Eccles (1953) has proposed this phenomenon of posttetanic potentiation as a general model for conditioning and memory. Such a proposal meets with many difficulties (Malmo, 1954). However, there is no question that a person ascribing learning to changes in synaptic excitability could do so with more confidence today than was possible 30 years ago.

Conclusions

In the opinion of the writer, the over-all weight of evidence certainly favors the existence of some mechanism of consolidation (in spite of the fact that alternative explanations are possible for many of the experiments which supposedly support the existence of such a process). Furthermore, the application of available physiological procedures appears to offer a promising approach to defining the structures involved in the fixation of memory traces. The most severe problems presented thus far have occurred as the result of confounds in the behavioral test situations employed, rather than through some defect in the modes of physiological interference. These problems are not insoluble, however, and an attempt was made to indicate this in the text of the paper.

As a final point, the material reviewed suggests the possibility that pseudoneurological speculation, resulting from strictly behavioral observation, can result in productive physiological research—when the speculation is shrewdly conceived. Moreover, the physiologist would appear to have already begun to repay this debt by suggesting purely behavioral studies or new interpretations of behavioral data. The studies demonstrating interfering effects of visual stimulation interpolated immediately after visual discrimination learning (Thompson, 1957b; Thompson & Bryant, 1955) are examples of such physiologically influenced "behavioral" investigations. Along similar lines, Walker's (1958) reinterpretation of reaction decrement, spontaneous alternation data, in terms of mechanisms serving to protect consolidation, appears to be equally sensitive to current physiological research.

REFERENCES

Adey, W. R., Merrillees, N. C. R., & Sunderland, S. The entorhinal area: Behavioural, evoked potential, and histological studies of its interrelationships with brain-stem regions. *Brain*, 1956, **79**, 414–439.

Ballard, P. B. Oblivescence and reminiscence. *Brit. J. Psychol., Monogr. Suppl.*, 1913, **1**, No. 2.

Berkun, M. M., Kessen, Marion L., & Miller, N. E. Hunger-reducing effects of food by stomach fistula versus food by mouth measured by a consummatory response. *J. comp. physiol. Psychol.*, 1952, **45**, 550–554.

Boycott, B. B., & Young, J. Z. The comparative study of learning. *Symp. Soc. Exp. Biol.*, 1950, **4**, 432.

Brady, J. V. The effect of electroconvulsive shock on a conditioned emotional response: The permanence of the effect. *J. comp. physiol. Psychol.*, 1951, **44**, 507–511.

Brady, J. V. The effect of electroconvulsive shock on a conditioned emotional response: The significance of the interval between the emotional conditioning and the electroconvulsive shock. *J. comp. physiol. Psychol.*, 1952, **45**, 9–13.

Brady, J. V., Boren, J. J., Conrad, D., & Sidman, M. The effect of food and water deprivation upon intracranial self-stimulation. *J. comp. physiol. Psychol.*, 1957, **50**, 134–137.

Brady, J. V., Schreiner, L., Geller, I., & Kling, A. Subcortical mechanisms in emotional behavior: The effect of rhinencephalic injury upon the acquisition and retention of a conditioned avoidance response in cats. *J. comp. physiol. Psychol.*, 1954, **47**, 179–186.

Burnham, W. H. Retroactive amnesia: Illustrative cases and a tentative explanation. *Amer. J. Psychol.*, 1903, **14**, 382–396.

Burns, B. D. The production of afterbursts in isolated unanesthetized cerebral cortex. *J. Physiol.*, 1954, **125**, 427–446.

Burns, B. D. *The mammalian cerebral cortex.* London: Arnold, 1958.

Burns, N. M., & Mogenson, G. Effects of cortical stimulation on habit acquisition. *Canad. J. Psychol.*, 1958, **12**, 77–82.

Burns, N. M., & Stackhouse, S. P. Effects of cortical stimulation on habit acquisition: Confirmatory data. *Amer. Psychologist*, 1959, **14**, 427. (Abstract)

Carson, R. C. The effect of electroconvulsive shock on a learned avoidance response. *J. comp. physiol. Psychol.*, 1957, **50**, 125–129.

Cerf, J. A., & Otis, L. S. Heat narcosis and its effect on retention of a learned behavior in the goldfish. *Federation Proc.*, 1957, **16**, 20–21. (Abstract)

Coons, E. E., & Miller, N. E. Conflict versus consolidation of memory traces to explain "retrograde amnesia" produced by ECS. *J. comp. physiol. Psychol.*, 1960, **53**, 524–531.

Cronholm, B., & Molander, L. Influence of an interpolated ECS on retention of memory material. *U. Stockholm Psychol. Lab. Rep.*, 1958, No. 61.

DeCamp, J. E. A study of retroactive inhibition. *Psychol. Monogr.*, 1915, **19**(4, Whole No. 84).

Duncan, C. P. The retroactive effect of electroshock on learning. *J. comp. physiol. Psychol.*, 1949, **42**, 32–44.

Eccles, J. C. *The neurophysiological basis of mind.* Oxford: Clarendon, 1953.

Eccles, J. C. *The physiology of nerve cells.* Baltimore: Johns Hopkins, 1957.

Essman, W. B., & Jarvik, M. E. The retrograde effect of ether anesthesia on a conditioned avoidance response in mice. *Amer. Psychologist*, 1960, **15**, 498. (Abstract)

Fay, T. Observations on prolonged human refrigeration. *NY State J. Med.*, 1940, **40**, 1351–1354.

Fisher, A. E. Maternal and sexual behavior induced by intracranial stimulation. *Science*, 1956, **124**, 228–229.

Flescher, D. L'amnesia retrograda dopo l'ettroshock: Contributo allo studio della patogenesi delle amnesia in genere. *Schweiz. Arch. Neurol. Psychiat.*, 1941, **48**, 1–28.

French, J. W. The effect of temperature on the retention of a maze habit in fish. *J. exp. Psychol.*, 1942, **31**, 79–87.

Gallinek, A. Fear and anxiety in the course of electroshock therapy. *Amer. J. Psychiat.*, 1956, **113**, 428–434.

Geller, I., Sidman, M., & Brady, J. V. The effect of electroconvulsive shock on a conditioned emotional response: A control for acquisition recency. *J. comp. physiol. Psychol.*, 1955, **48**, 130–131.

Gerard, R. W. Biological root of psychiatry. *Science*, 1955, **122**, 225–230.

Glickman, S. E. Deficits in avoidance learning produced by stimulation of the ascending reticular formation. *Canad. J. Psychol.*, 1958, **12**, 97–102.

Glickman, S. E. Reinforcing properties of arousal. *J. comp. physiol. Psychol.*, 1960, **53**, 68–71.

Green, J. D., & Adey, W. R. Electrophysiological studies of hippocampal connections and excitability. *EEG clin. Neurophysiol.*, 1956, **8**, 245–262.

Hayes, K. J. Anoxic and convulsive amnesia in rats. *J. comp. physiol. Psychol.*, 1953, **46**, 216–217.

Hebb, D. O. *The organization of behavior.* New York: Wiley, 1949.

Hebb, D. O. The role of neurological ideas in psychology. *J. Pers.*, 1951, **20**, 39–55.

Hudson, B. B. One-trial learning in the domestic rat. *Genet. psychol. Monogr.*, 1950, **41**, 99–145.

Hunter, W. S. The effect of inactivity produced by cold upon learning and retention in the cockroach. *Blatella Germanica. J. genet. Psychol.*, 1932, **41**, 253–266.

Jones, M. R. The effect of hypothermia on retention. *J. comp. physiol. Psychol.*, 1943, **35**, 311–316.

Kohn, H. I. Changes in blood of the rat during ether and barbiturate anaesthesia. *Amer. J. Physiol.*, 1950, **160**, 277–284.

Kohn, M. Satiation of hunger from food injected directly into the stomach versus food injected by mouth. *J. comp. physiol. Psychol.*, 1951, **44**, 412–422.

Lashley, K. S. A simple maze: With data on the relation of the distribution of practice to the rate of learning. *Psychobiology*, 1918, **1**, 353–367.

Lashley, K. S. *Brain mechanisms and intelligence.* Chicago: Univer. Chicago Press, 1929.

Leukel, F. P. The effect of ECS and pentothal anaesthesia on maze learning and retention. *J. comp. physiol. Psychol.*, 1957, **50**, 300–306.

Lloyd, D. P. C. Post-tetanic saturation of response in monosynaptic reflex pathways of the spinal cord. *J. gen. Physiol.*, 1949, **33**, 147–170.

Lorente de No, R. Analysis of the activity of the chains of internuncial neurons. *J. Neurophysiol.*, 1938, **1**, 207–244.

McDougall, W. Experimentelle Beiträge zur Lehre von Gedächtniss: Von G. E. Müller und A. Pilzecker. *Mind*, 1901, **10**, 388–394.

McGeoch, J. A. & Irion, A. L. *The psychology of human learning.* (2nd ed.) New York: Longmans, Green, 1952.

Magoun, H. W. *The waking brain.* Springfield, Ill.: Charles C Thomas, 1958.

Mahut. Helen. Discussion. In H. H. Jasper, L. D. Proctor, R. S. Knighton, W. C. Noshay, & R. T. Costello (Eds.), *Reticular formation of the brain.* Boston: Little, Brown, 1958.

Malmo, R. B. Eccles' neurophysiological model of the conditioned reflex. *Canad. J. Psychol.*, 1954, **8**, 125–129.

Miller, N. E., & Coons, E. E. Conflict versus consolidation of memory to explain "retrograde amnesia" produced by ECS. *Amer. Psychologist*, 1955, **10**, 394–395. (Abstract)

Milner, Brenda, & Penfield, W. The effect of hippocampal lesions on recent memory. *Trans. Amer. Neurol. Ass.*, 1955, 42–48.

Müller, G. E., & Pilzecker, A. Experimentelle Beiträge zur Lehre vom Gedächtnis. *Z. Psychol.*, 1900, Suppl. No. 1.

Olds, J., & Milner, P. Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *J. comp. physiol. Psychol.*, 1954, **47**, 419–427.

Olds, J., & Olds, M. E. Positive reinforcement produced by stimulating hypothalamus with iproniazid and other compounds. *Science*, 1958, **127**, 1175–1176.

Olds, J., & Peretz, B. Relation of arousal and motivational effects in reticular activating system. *Federation Proc.*, 1959, **18**, 116. (Abstract)

Osgood, C. E. *Method and theory in experimental psychology.* New York: Oxford Univer. Press, 1953.

Palay, S. L. Synapses in the central nervous system. *J. biophysic. biochem. Cytol., Suppl.*, 1956, **2**, 193–202.

Pearlman, C. A., Jr., Sharpless, S. K., & Jarvik, M. E. Retrograde amnesia produced by anesthetic and convulsant agents. *J. comp. physiol. Psychol.*, 1961, **54**, 109–112.

Pennington, D. F. The effects of ECS on retention of a discrimination habit in rats subjected to anoxia. *J. comp. physiol. Psychol.*, 1958, **51**, 687–689.

Pillsbury, W. B. *The essentials of psychology.* New York: Macmillan, 1913.

Poschel, P. B. H. Proactive and retroactive effects of electroconvulsive shock on approach-avoidance conflict. *J. comp. physiol. Psychol.*, 1957, **50**, 392–396.

Ransmeier, R. E. The effects of convulsion, hypoxia, hypothermia and anesthesia on retention in the master. Unpublished doctoral dissertation, University of Chicago, 1953.

Ransmeier, R. E., & Gerard, R. W. Effects of temperature, convulsion and metabolic factor on rodent memory and EEG. *Amer. J. Physiol.*, 1954, **179**, 663–664. (Abstract)

Rapaport, D. *Emotions and memory.* Baltimore: Williams & Wilkins, 1942.

Ribot, T. Memory. In D. H. Tuke (Ed.), *A dictionary of psychological medicine.* Philadelphia: Blakiston, 1892.

Rosvold, H. E., & Delgado, J. M. R. The effect on delayed-alternation test performance of stimulating or destroying electrically structures within the frontal lobes of the monkey's brain. *J. comp. physiol. Psychol.*, 1956, **49**, 365–372.

Russell, R. W., & Hunter, W. S. The effects of inactivity produced by sodium amytal on the retention of the maze habit in the albino rat. *J. exp. Psychol.*, 1937, **20**, 426–436.

RUSSELL, W. R., & NATHAN, P. W. Traumatic amnesia. *Brain*, 1946, **69**, 280–300.

SCOVILLE, W. B., & MILNER, BRENDA. Loss of recent memory after bilateral hippocampal lesions. *J. Neurol. Neurosurg. Psychiat.*, 1957, **20**, 11–21.

SHERRINGTON, C. S. *The integrative action of the nervous system.* New York: Scribner, 1906.

STAINBROOK, E. Shock therapy: Psychologic theory and research. *Psychol. Bull.*, 1946, **43**, 21–60.

STAINBROOK, E. The effects of electrically induced convulsions on animal behavior. *J. Pers.*, 1948, **17**, 2–8.

STELLAR, E. Physiological psychology. *Annu. Rev. Psychol.*, 1957, **8**, 415–436.

THOMPSON, R. The comparative effects of ECS and anoxia on memory. *J. comp. physiol. Psychol.*, 1957, **50**, 397–400. (a)

THOMPSON, R. Retroactive effect of interpolated visual stimulation. *Psychol. Rep.*, 1957, **3**, 183–188. (b)

THOMPSON, R. The effect of ECS on retention in young and adult rats. *J. comp. physiol. Psychol.*, 1958, **50**, 644–646. (a)

THOMPSON, R. The effect of intracranial stimulation on memory in cats. *J. comp. physiol. Psychol.*, 1958, **51**, 421–426. (b)

THOMPSON, R., & BRYANT, J. H. Memory as affected by activity of the relevant receptor. *Psychol. Rep.*, 1955, **1**, 393–400.

THOMPSON, R., & DEAN, W. A further study on the retroactive effect of ECS. *J. comp. physiol. Psychol.*, 1955, **48**, 488–491.

THOMPSON, R., HARAVEY, F., PENNINGTON, D. F., SMITH, J., JR., GANNON, D., & STOCKWELL, F. An analysis of the differential effects on ECS on memory in young and adult rats. *Canad. J. Psychol.*, 1958, **12**, 83–96.

THOMPSON, R., & PENNINGTON, D. F. Memory decrement produced by ECS as a function of the distribution of original learning. *J. comp. physiol. Psychol.*, 1957, **50**, 401–404.

THOMPSON, R., & PRYER, R. S. The effect of anoxia on the retention of a discrimination habit. *J. comp. physiol. Psychol.*, 1956, **49**, 297–300.

WALKER, E. L. Action decrement and its relation to learning. *Psychol. Rev.*, 1958, **65**, 129–142.

WILLIAMS, M. Memory studies in electric convulsion therapy. *J. Neurol. Neurosurg. Psychiat.*, 1950, **13**, 30–35.

WOODWORTH, R. S. *Experimental psychology.* New York: Holt, 1938.

WORCHEL, P., & GENTRY, G. Electroconvulsive shock and memory: The effect of shocks administered in rapid succession. *Comp. psychol. Monogr.*, 1950, **20**, No. 104, 95–119.

YOUNG, J. Z. Discrimination and learning in octopus. In H. Von Foerster (Ed.), *Cybernetics: Transactions of the ninth conference.* New York: Josiah Macy, Jr. Foundation, 1953.

ZEIGLER, H. P. Electrical stimulation of the brain and the psychophysiology of learning and motivation. *Psychol. Bull.*, 1957, **54**, 363–382.

ZUBIN, J., & BARRERA, S. E. Effect of electric convulsive therapy on memory. *Proc. Soc. Exp. Biol. Med.*, 1941, **48**, 596–597. (Abstract)

(Received February 5, 1960)

Psychological Bulletin
1961, Vol. 58, No. 3, 234-255

THE ASSESSMENT OF ANXIETY BY PHYSIOLOGICAL BEHAVIORAL MEASURES[1]

BARCLAY MARTIN
University of Wisconsin

The term anxiety has enjoyed great popularity in the writings and researches of psychologists in the last decade, and procedures for measuring this hypothetical state have proliferated wildly. There is every indication that psychologists will continue to develop and employ measures of anxiety in many areas of research, especially in the rapidly expanding number of studies of psychotherapeutic process and change, in the already booming area of psychopharmacology, in studying the effects of anxiety on performance, and in attempts to assess such constructs as aggression anxiety or sex anxiety. It is the purpose of this paper to first impose some restrictions upon the definition of anxiety, and then to focus upon the problem of assessment by physiological-behavioral measures. No attempt will be made in this paper to review the research and evaluate the problems associated with assessing anxiety by self-report techniques.

One's theoretical approach to anxiety affects how one goes about measuring it; likewise results of attempts to assess anxiety should eventually modify and help refine the theoretical conception of anxiety. Thus the initial comments about the nature of anxiety should be considered as a rough formulation only, with both assessment procedures and theory modifying each other as investigation proceeds. It is recognized that this formulation, rough as it is, cannot include all that anxiety means to all people, and that accordingly to make this review manageable, it is necessary to delimit the concept.

A Conception of Anxiety

As a starting point it is proposed that the construct of anxiety be considered similar and perhaps identical to the reaction of fear, the neurophysiological bases for which are not completely known but would seem to especially involve the functions of the posterior hypothalamus and its effects upon the sympathetic nervous system, the adrenal medulla, and the pituitary-adrenocortical system. The brain stem reticular formation may also play a part in this reaction. It is recognized that this is undoubtedly an oversimplification of the complex and interacting neurophysiological mechanisms involved in fear. This reaction may be largely innate yet it is likely that as a result of learning or constitutional predisposition individuals tend to have variations in the manner in which the anxiety reaction is expressed.

It is further proposed that anxiety represents only one of many arousal states that can be differentiated from a more general state of activation as arousal becomes more intense. Thus the arousal that occurs when a person passes from a sleeping or very relaxed state to a waking, behaving state may be of a fairly generalized sort with no specialized affective or motivational reactions involved. However, as arousal becomes more intense, differentiation probably occurs and distinctive arousal states may emerge relating to such constructs as anxiety,

[1] This paper was prepared in part while the author held a visiting appointment at the University of North Carolina. The author wishes to thank Earl Baughman and Leonard Berkowitz for their helpful suggestions.

anger, hunger, sex, or other emotional or motivational states. Although it is possible that research will suggest the value of distinguishing anxiety from fear at the response level, or one kind of anxiety from another, it is perhaps best to demonstrate the utility of one construct of anxiety and its distinctiveness from other arousal states before adding unnecessarily to the number of theoretical constructs extant.

Anxiety also possesses the property of being highly learnable: that is, the hypothetical response becomes readily conditioned to stimuli that do not innately elicit the response. This characteristic renders difficult if not impossible any attempt to define anxiety on the basis of stimuli that elicit it, since the stimuli that elicit it will vary widely from person to person. An exception would be direct electrical stimulation of the brain (Miller, 1958), where the effective antecedent stimulus might be well defined.

As a consequence of the difficulty of approaching the construct of anxiety from the stimulus side in human subjects, the primary emphasis in this paper will be to review research relevant to the assessment of anxiety in terms of response patterns. The observable responses from which one might infer the strength of the anxiety reaction are of two basic types: physiological-behavioral responses and self-report responses. As previously mentioned, this paper will be primarily concerned with the first type of response.

In addition to the hypothetical anxiety state and its observable manifestations there are two other variables intimately related to anxiety which are kept conceptually distinct in the present view: namely, those stimuli (external or internal) which elicit the anxiety response, and those responses which have been learned because they reduce or avoid the anxiety response. From the point of view of measurement the stimuli that evoke anxiety become important only if one wants to know what situations or thoughts or feelings elicit anxiety. Thus the common distinction between anxiety and fear in terms of the latter being in response to a realistic danger and the former being a response to unrealistic or unknown threats is basically a stimulus defined difference and does not necessarily involve a difference in response.

There exists a possible source of confusion with respect to the responses that have been learned to reduce anxiety in that clinicians frequently infer anxiety on the basis of these "defenses" against anxiety as much as from direct expression of the anxiety itself. Again, from the point of view of theory as well as measurement it is preferable to keep these two variables distinct if possible. In fact, it would seem likely that when a person is making a successful "defensive" response, no anxiety is present. To the extent that this is so it would be misleading to infer the strength of the momentary anxiety level from the presence of learned anxiety reducing responses.

The Measurement of Anxiety

The foregoing theoretical analysis suggests that in spite of individual variations in response there might still be some pattern of physiological-behavioral responses associated with anxiety arousal that would be distinct from other patterns of response associated with other emotional or arousal states. Findings based primarily on physiological response patterns will be considered first followed by findings based primarily on behavioral response patterns. Two basic questions will be asked with respect to both the physiological and the behavioral evidence: (*a*) Does a distinctive pattern of responses emerge, tentatively identifiable as reflecting anxiety, that can be distinguished from other patterns associated with other arousal states, when the differing arousal states have been experimentally induced? (*b*) What is the nature of the intercorrelations among physiological or behavioral measures which have been obtained under the same experimental conditions, and is there any evidence of a distinguishable cluster of intercorrelated variables that might be tentatively identified as reflecting anxiety? The studies do not always lend themselves to a clear-cut analysis in these terms but these are the guiding questions being considered.

Physiological Measures: Experimental Comparisons

The studies of primary interest here are those in which an attempt was made to distinguish between two or more experimentally induced arousal states where one of these was considered to represent a fear or anxiety reaction. There are three studies that most closely follow this paradigm. Ax (1953) reports a study in which a variety of physiological measures were obtained from normals under conditions presented in counterbalanced order that were designed to elicit fear and anger, respectively. The fear condition was ingeniously contrived to make the subject think that the apparatus was faulty and that he was in real danger of receiving a severe, perhaps even fatal, electric shock. Anger was aroused by an obnoxious assistant who generally insulted and belittled the subject. Schachter (1957) repeated Ax' study using hypertensive, potential hypertensive, and normotensive subjects, and added a pain experience (cold pressor test) to the fear and anger situations. All subjects received the treatments in the same order: pain, fear, and anger. Lewinsohn (1956) obtained three physiological measures plus a measure of finger tremor on groups of normals, anxiety reaction patients, ulcer patients, and hypertensive patients subjected in counterbalanced order to the cold pressor test and a failure experience accompanied by criticism and electric shock. Another study that is highly relevant to the issue but which employed a somewhat different research strategy is that of Funkenstein, King, and Drolette (1957). After stressing their college student subjects they determined in a poststress interview whether a subject had tended to experience anger outwardly directed, anger inwardly directed, or anxiety. The scores obtained were limited to blood pressure and ballistocardiographic measures.

The results of these four studies are summarized in Table 1. Most scores in the Ax and Schachter studies represent difference scores between prestress resting level and the highest (or in some cases the lowest) level reached during stress. The scores in the Lewinsohn study represent differences in the mean during rest and the mean during stress, with the exception of the GSR score which represents the largest deflection during stress. All scores reported from the Funkenstein study are percentage changes from prestress levels.

In spite of some inconsistencies among the studies there does appear to be evidence for distinguishable response patterns that can be tentatively associated with the constructs of fear (anxiety) and anger. Diastolic blood pressure increased more for anger than fear in all three studies in which fear and anger states were thought to be aroused (significantly different from chance in two studies). Heart rate increased more in fear than anger in all three studies (significant in two). Maximum heart rate

TABLE 1

COMPARISON OF PHYSIOLOGICAL MEASURES ASSOCIATED WITH DIFFERENT EMOTIONAL AROUSAL STATES IN FOUR STUDIES

Measure	Ax		Schachter			Lewinsohn		Funkenstein	
	Fear	Anger	Fear	Anger	Pain	Fear	Pain	Fear	Anger-out
Systolic blood pressure	20.4	19.2	22.5	21.1	17.8			19.6%*	13.1%*
Diastolic blood pressure	14.5*	17.8*	13.7	14.5	11.8			9.7%	22.8%
Heart rate (+)	30.3	25.8	18.7*	10.8*	0.3*	5.4*	0.9*	33.3%*	7.4%*
Heart rate (−)	4.0*	6.0*							
Cardiac output			6.7*	3.0*	−0.25*			61.9%*	−3.2%*
Peripheral resistance			−1.10*	0.04*	1.28*			−19.3%*	32.9%*
Hand temperature (−)	.045	.050	.036*	.030*	.024*				
Palmar conductance	14.8*	9.4*	−1.99[a]*	−2.18[a]*	−2.33[a]*				
Largest deflection in stress, GSR						2.52	2.15		
No. GSRs	4.7*	11.6*							
Respiratory rate	6.0*	2.3*	2.8*	2.1*	0.7*				
Frontalis muscle tension	3.34*	4.35*	1.30	2.26	1.65				
No. muscle potential peaks	13.2*	10.5*							
Finger tremor						87	118		
Salivary output						−0.9	7.9		

[a] Schachter used the transformation, log $1/(R_1-R_2)$, where R_1 = initial resistance and R_2 = lowest resistance during stress. The smallest negative number, −1.99, for fear accordingly refers to the largest decrease in resistance.

* Significant at the .05 level; for Schachter this is based on an overall analysis of variance for the three conditions.

decrease was significantly greater in anger than fear in the one study in which it was reported. Cardiac output increased significantly more in fear than anger in the two studies in which it was reported, and peripheral resistance decreased significantly more in fear than anger in both studies where it was reported. Palmar conductance increased significantly more in fear than anger in the two studies where it was reported. Number of discrete GSRs, however, was significantly higher in anger than fear in the one study where this was measured. Respiration rate increased significantly more in fear than anger in the two studies reporting this measure. Frontalis muscle tension increased more in fear than anger in the two studies measuring it (significant in one).

Another study that was not included in the tabular presentation provides additional support for the different heart rate responses associated with anxiety and anger. DiMascio, Boyd, and Greenblatt (1957) studied one psychotherapy patient over 11 interviews and found a correlation (rho) of .69 between average heart rate and amount of rated tension (anxiety?) in the interviews, and a correlation of −.37 between average heart rate and amount of rated antagonism in the interviews.

The two studies in Table 1 involving a painful experience, the cold pressor test, suggest that this arousal state may also be distinguishable from fear, although the differentiation of pain and anger is less clear. It is, of course, not possible to know from these results how specific these reactions might be to the cold pressor test as opposed to pain stimulation generally.

Funkenstein et al. (1957) propose a theory that may serve to provide some integration for these various findings. They suggest that the physiological reaction accompanying anger-out is a norepinephrine-like reaction and that accompanying anxiety is an epinephrine-like reaction. The physiological reactions accompanying injections of epinephrine and norepinephrine have been investigated by Goldenberg, Pines, Baldwin, Greene, and Roh (1948), Barcroft and Konzett (1949), DeLargy, Greenfield, McCorry, and Whelan (1950), Goldenberg (1951), Swan (1952), and Clemens (1957). In general it is found that epinephrine leads to increased palmar conductance, systolic blood pressure, heart rate, cardiac output, forehead temperature, central nervous system stimulation, blood sugar level; and decreased diastolic blood pressure, peripheral resistance, hand temperature, and salivary output. Norepinephrine leads to increased systolic and diastolic blood pressure and peripheral resistance, no change or a slight decrease in heart rate and cardiac output, and only slight increases in central nervous system stimulation and blood sugar level.

It is generally thought that reactions associated with norepinephrine are more limited, possibly restricted to peripheral vasoconstriction resulting from secretion at the sympathetic nerve endings, than are the reactions to epinephrine. However, no studies were found in which the effects of injected norepinephrine upon a wide range of responses including palmar conductance, hand or finger temperature, respiration rate, salivary output, or muscle potentials were assessed. In terms of the measures that have been obtained under both kinds of hormonal injections (Barcroft & Konzett, 1949; DeLargy et al., 1950; Goldenberg et al., 1948), heart rate, diastolic blood pressure, cardiac output, and peripheral resistance appear to be the most discriminating. Neither cardiac output nor peripheral resistance is readily obtainable by direct measurement. Cardiac output is usually inferred from ballistocardiographic measures, and peripheral resistance is usually estimated by dividing mean arterial blood pressure by cardiac output.

Funkenstein et al. (1957) divided their subjects into subgroups on the basis of epinephrine-like, norepinephrine-like, and indeterminate reactions and found a highly significant relationship in the expected direction between these physiological reaction types and the tendency to respond by anger-out as opposed to anxiety. Schachter (1957) making use of a greater variety of physiological measures likewise computed an index of epinephrine- and norepinephrine-like reactions and found these indices to vary significantly as a function of the pain, anger, and fear conditions with pain showing the most norepinephrine-like reaction and fear the most epinephrine-like reaction with anger falling in between.

Although it would be premature to conceptualize the anxiety reaction as being entirely defined by the results of epinephrine secretion, the distinction between the epinephrine- and norepinephrine-like reactions may well be an important one for anxiety measurement. The secretion of epinephrine and norepinephrine from the adrenal medulla and the release of norepinephrine at the sympathetic nerve endings are all affected by sympathetic nervous system stimulation. The fact that these two hormones produce quite different reactions points up what has long been known: namely, that it is a great oversimplification to speak of sympathetic arousal as if it were a unitary function. Although the response pattern associated with experimentally induced anxiety conforms rather closely to the response pattern associated with epinephrine injection, the response pattern associated with anger is not as closely related to the responses produced by norepinephrine injection. Perhaps the distinction between anxiety and anger, at the humoral level, is one involving the relation of epinephrine to norepinephrine in which anxiety is associated with a purer epinephrine-like reaction and anger with a mixed pattern of epinephrine and norepinephrine responses.

There are other studies where one or two physiological measures have been obtained under conditions likely to arouse anxiety. For example, Hickham, Cargill, and Golden (1948) found heart rate and cardiac output to increase substantially in medical students before what was considered to be an anxiety arousing situation, an oral examination, as compared to more relaxed conditions a month later. Likewise, Malmo, Boag, and Smith (1957) report increased heart rate in neurotic subjects after criticism as compared with decreased heart rate after praise. Although studies of this kind tend to be consistent with the previously described studies, they do not shed additional light on the question of whether some pattern of response related to anxiety can be differentiated from patterns of response associated with other kinds of arousal states.

Davis (1957), Davis and Buchwald (1957), and Davis, Buchwald, and Frankman (1955) also report evidence that different stimuli elicit distinctive autonomic response patterns. There is no reason to believe, however, that any of their stimuli, for example, pictures of nudes, landscapes, etc., were likely to evoke anxiety in many of their subjects. These studies do point to the possible subtleties in autonomic patterns associated with various kinds of stimulation or arousal states, and caution against any too ready acceptance of some particular pattern as being *the* anxiety or *the* anger pattern. All of the studies described thus far, though, are consistent with the possibility that some pattern of physiological measures may allow one to infer the magnitude of the hypothetical anxiety reaction differentially from other hypothetical states such as anger or pain.

Physiological Measures: Group Comparisons

There is a host of studies in which physiological measures are contrasted between normals and various clinical groups presumed to be in general more anxious than the normals. The studies that will be considered here are those involving patient groups in which the presence of manifest anxiety was reported to be a prominent part of the symptom picture; accordingly, much of the physiological research on such psychosomatic disorders as hypertension and peptic ulcer will not be summarized.

Sherman and Jost (1942) found 15 neurotic children to have *lower* resting level palmar conductance than 18 well adjusted children, but more resting level hand tremors, lower percentage of alpha rhythm in the EEG, and faster respiration rate than well adjusted children. No differences were found for heart rate or blood pressure. Although measures were taken in a series of seven conditions, the results described above appeared to represent differences in general level rather than different degrees of reaction to the various conditions. Jurko, Jost, and Hill (1952) obtained measures on 25 normals, 20 neurotics, and 10 schizophrenics (all adults) while administering the Rosenzweig P–F test, and found heart rate, respiration rate, and respiration variability higher in patient than normal groups before and during the test administration. A body movement score was highest for the schizophrenics and lowest for the normals. Palmar conductance was again found to be inconsistent with the general pattern, being highest for the normals and lowest for the schizophrenics before and during test administration. In neither of these two studies, however, was any attempt made to restrict the sample of neurotics to patients in which anxiety was the most prominent symptom.

GSR conditioning rate, on the other hand, has been found to be faster in more anxious subjects (Bitterman & Holtzman, 1952; Schiff, Dougan, & Welch, 1949; Welch & Kubis, 1947).

White and Gildea (1937) found that patients in which anxiety was a prominent symptom showed greater heart rate increases to the cold pressor test than did normals. On the surface such a finding appears contradictory to the results of Schachter (1957) in which the physiological responses associated with the cold pressor test were clearly distinguishable from those associated with anxiety. White and Gildea, however, obtained measures during a rest period, during a brief anticipation period in which the experimenter moved the dish of ice water close to the subject, and during the immersion itself. For the normal group the average heart rates for these three periods were 75.7, 81.5, and 80.0, respectively; and for a group of anxiety neurotics 81.0, 90.0, and 95.5, respectively. Clearly, it was the anticipation of the experience that led to increased heart rate for the normals, not the pain experience itself. The anxious patients likewise showed their greatest increase during anticipation. These results suggest that anticipation of the cold pressor test is anxiety arousing, and might yield a different pattern of response, in normals at any rate, than the pain experience itself.

The above results of White and Gildea as well as the results of Schachter (1957) and Lewinsohn (1956) argue against the theoretical formulation of Mowrer (1939) that anxiety (fear) is the conditioned form of the pain reaction.

In the Lewinsohn (1956) study previously mentioned, resting level palmar conductance was highest for the anxiety reaction group and lowest for the ulcer group, with normals and hypertensives falling in between. Resting level salivary output was highest in the ulcer group, and lower and about the same for the other three groups. Somewhat surprisingly, resting level heart rate was lowest for the anxiety group. The change scores showed no particular tendency to be associated with the diagnostic groups. Wishner (1953) found resting level heart rate to be higher in 11 anxiety neurotics than in 10 normals and a tendency, not significant, for respiration rate to be faster in the neurotics. Funkenstein, Greenblatt, and Solomon (1951, 1952) conclude that patients with anxiety and depressive symptoms are manifesting a chronic epinephrine-like reaction, whereas patients with paranoid tendencies or who are otherwise directing their anger and blame upon the external world are manifesting a chronic norepinephrine-like reaction. Their conclusions are based primarily on the patients' reactions to the mecholyl test (Funkenstein, Greenblatt, & Solomon, 1950).

Malmo (1950, 1957) has summarized his research with respect to physiological measures found to discriminate between normals and patients with pathological degrees of anxiety. In his 1957 article he concludes that anxious patients show greater reactivity in many measures regardless of the kind of stress used. Thus, Malmo and Shagass (1949a) using a painful thermal stimulation of the forehead as their stress found anxiety neurotics and early schizophrenics to show more finger movements, greater neck muscle potentials, more head movements, more respiratory irregularities, and greater heart rate variability than normal controls. Percent change of the GSR showed no significant relationship. These results have been generally borne out in other studies using different stresses: Malmo, Shagass, and Davis (1951); Malmo, Shagass, Belanger, and Smith (1951). The results of Malmo and Smith (1955) suggest frontalis muscle tension may be a more sensitive discriminator between normals and anxiety neurotics than forearm muscle tension.

Wenger (1948) using considerably larger *N*s than most investigators compared resting state physiological measures of 225 patients with the diagnosis of operational fatigue, 98 hospitalized psychoneurotics, and a normative group of 488 unselected preflight students in the Army Air Force. The 10 measures that significantly discriminated between the operational fatigue group and the normal group were salivary output, palmar conductance, systolic and diastolic blood pressures, sinus arrhythmia, heart period, sublingual temperature, finger temperature, respiration period, and tidal air mean. The operational fatigue group showed sympathetic dominance on all of the above measures except sublingual temperature. For 47 patients in the operational fatigue group Wenger obtained repeat measures on most variables at a later time when they were considered improved and ready to return to duty. Of the 20 variables tested only palmar conductance, heart period, and finger temperature showed significant changes, and these were all in the direction of lessened sympathetic arousal. The results with respect to the hospitalized psychoneurotics, although not yielding exact correspondence on specific measures, also showed a strong sympathetic dominance for this clinical group.

Gunderson (1953) obtained 12 resting state autonomic measures, selected on the basis of Wenger's previous work, on a sample of 110 early schizophrenics with an average length of hospitalization of about 2 years. Nine measures—salivary output, dermographic latency, dermographic persistence, systolic blood pressure, diastolic blood pressure, finger temperature, heart rate, respiration rate, and sublingual temperature—were significantly different from Wenger's normative group of aviation cadets, and with the exception of sublingual temperature all were in the direction of greater sympathetic arousal. Palmar conductance failed to discriminate and was, in fact, almost identical for the two groups. This schizophrenic sample also showed significantly greater sympathetic arousal in seven of these measures than Wenger's neurotic group. As Gunderson points out this indication of greater anxiety in the schizophrenic group may well not exist in more chronic patients. Gunderson also divided the schizophrenic subjects into those that had improved the most and least with shock therapy and found the most improved group to show less general sympathetic arousal as measured by Wenger's autonomic balance score, the conclusion being that improvement had been accompanied by a decreased arousal.

There are difficulties involved in comparing these studies in which anxiety is assumed to be present by virtue of a psychiatric diagnosis with those in which anxiety was produced experimentally. For example, if anger or annoyance does involve a distinctive arousal state and if such a state is present more often in some of these patient groups than in normals, a not unlikely assumption, then the pattern of mean scores may reflect a mixture of anxiety and anger as well as other arousal states. Nevertheless, many measures which belong to the epinephrine-like pattern of reaction are found to consistently discriminate, with an occasional exception, between anxious patients and normals. By and large it would appear that so-called resting state measures discriminate between the patients and normals as well, and in some cases better, than do change scores associated with experimental stress. Some of the studies reporting change score results may be misleading since in most cases the patient groups start out with higher initial level scores. The high negative correlation between initial level and the magnitude of the change score that prevails for most autonomic measures might well obscure some real differences that would have emerged if this correlation had been partialed out by some procedure such as Lacey's (1956) autonomic lability score.

It is also possible that the particular pattern of autonomic responses associated with an immediate threat situation is different from the "steady state" pattern of more chronically elevated responses found in many psychiatric patients. It is interesting in this regard that Wenger (1957) in recent pattern analyses of the data in his various samples reports not only patterns of sympathetic and parasympathetic dominance but a pattern composed of a mixture of sympathetic and parasympathetic type of responses. This latter pattern, which Wenger calls the B pattern, consists of three sympathetic type tendencies, high heart rate, high systolic blood pressure, and low salivary output; and two characteristics of parasympathetic innervation or lack of sympathetic arousal, high finger temperature and low palmar conductance. The sympathetic pattern occurred more frequently in neurotic and schizophrenic samples than in the normal group, but not more frequently in the operational fatigue or a psychosomatic sample than in the normal group. The B pattern occurred more frequently in all of the four psychiatric groups than in the normal group. Perhaps this B pattern represents a more chronic result of psychological stress which could be distinguished from the anxiety state as presently conceived. Such an interpretation is consistent with the common clinical view that psychosomatic symptoms frequently serve an anxiety reducing function. It is also noteworthy that the findings of Sherman and Jost (1942) and Jurko, Jost, and Hill (1952) of low resting level palmar conductance in a pattern otherwise suggestive of sympathetic activation in neurotic patients is consistent with the existence of Wenger's B pattern.

To carry speculation a bit further in this area, it may be that there are systematic differences in response patterns as a function of the chronicity of the stress, as suggested by Selye (1950). Thus the pattern(s) of immediate change scores associated with discrete stimuli (electric shock or a threatening word) may be different from the pattern(s) of response associated with a stress of longer duration but still essentially temporary or situational (oral examination, the general situation in an electric shock experiment, or an appointment for a first psychotherapy hour), where the change scores would have to be based upon measures obtained at some more relaxed time. And both of the above kinds of patterns might differ from patterns of response resulting from stress continuing over months or years as would be the case with psychiatric patients. The distinctive characteristics of responses associated with the second as opposed to the first type of stress may result from humoral effects being added to the more direct and shorter latency effects of autonomic nervous system stimulation.

There have been several other approaches to the physiological assessment of anxiety employing measures less readily obtainable and also less amenable to continuous recording than most of the ones considered above. Ulett, Gleser, Winokur, and Lawler (1953) and Shagass (1955b) report that the EEG of anxious patients can be more readily "driven" at higher frequencies than is the case for normals or less anxious patients. There was no tendency for the average undriven alpha frequency to be different for the groups. Shagass (1955a) further reports that changes in the driven EEG frequency correspond to changes in anxiety level for the same person measured at different times.

Sedation threshold is also reported by Shagass (1954) and Shagass and Naiman (1955) to be related to anxiety level in patients. Basowitz, Persky, Korchin, and Grinker (1955) find more hippuric acid in the urine of paratrooper trainees assessed to be anxious than those not anxious, and also more in anxiety neurotics than in normals.

Physiological Measures: Intercorrelations

On the basis of the research just summarized one might assume that many of the measures found to be related to experimentally induced or clinically assessed anxiety would show substantial intercorrelations. Research thus far gives little ground for optimism that these variables will correlate very highly, if at all. However, it should be pointed out that there are few researches that provide much direct evidence on the question: namely, correlations among changes in measures obtained under resting and a clearly fear or anxiety arousing situation. Ax (1953) intercorrelated the seven physiological change scores that significantly discriminated between the fear and anger conditions. The intercorrelations of these scores under the anger condition tended to be higher than for the fear condition. The correlations were for the most part insignificant for fear. Schachter (1957) did not report intercorrelations among his measures but did find significantly more variability among the measures under fear than anger. Lewinsohn (1956) likewise reported intercorrelations among his four variables for base level scores, for change scores to the cold pressor test, and for change scores to the failure-criticism condition. Only a few correlations were significant, probably not more than could have occurred by chance. Terry (1953) intercorrelated a number of physiological change scores associated with doing arithmetic problems under distracting noise conditions. The intercorrelations between different autonomic systems were very low and for the most part insignificant. Only measures of closely related functions, such as systolic and diastolic blood pressure, correlated to any degree. It is possible that the stress condition was not particularly anxiety arousing for most subjects.

Sherman and Jost (1942) in contrast to the above studies did find a number of significant correlations among their physiological variables for neurotic and normal children combined. Although their correlation matrix is based on a mixture of absolute level scores, percent change scores, and scores obtained at different points in a sequence of seven conditions, there does seem to be a cluster of fairly highly intercorrelated variables suggesting some arousal dimension. The measures most highly intercorrelated are hand tremor, percent heart rate change, percent alpha dominance (negative correlations), and respiratory variability. Weybrew (1959) intercorrelated 12 physiological change scores and 4 personality ratings. The physiological measures were obtained before and after the subjects were subjected to a standardized situational stress. Correlations were in general low among the physiological change scores, and the results of a factor analysis were not easy to interpret.

There are just not enough studies with enough significant correlations between change scores to attempt any generalizations from the results. A general problem encountered in working with autonomic change scores is with respect to the type of transformation, if any, to use. Correlations, for example, among Lacey's (1956) autonomic lability scores would appear to provide a more meaningful picture of the tendency of measures to covary than would be obtained by using absolute change, percentage change, or most other transformations, since as previously mentioned Lacey's score more adequately partials out the usual high negative correlation between change and initial level. The degree to which correlations among autonomic change scores can be affected by partialing out the correlation with initial level is shown in the results of Mandler and Kremen (1958). They intercorrelated scores obtained under a failure stress condition from five different response systems (GSR, heart rate, respiration, face temperature, and blood volume) including in some cases absolute change scores along with Lacey's autonomic lability score. Absolute heart rate change yielded a correlation of .27 with change in respiration rate, whereas heart rate with initial level partialed out yielded a correlation of −.17; or in another case absolute heart rate change correlated only .02 with inspiration amplitude (with initial level of inspiration amplitude partialed out) but heart rate with initial level partialed out correlated .31 with the same measure. It is clear that correlations among autonomic measures will be greatly affected by the way in which the relation to initial level is handled.

The findings of Lacey (1950), Lacey and Van Lehn (1952), Lacey, Bateman, and Van Lehn (1953), even though based on stressors that for the most part cannot be accepted as clearly anxiety arousing, provide such a strong argument for individual patterns of autonomic response that they should not be ignored in this context. Using various samples (college students and mothers of children in the Fel's longitudinal research program) and various stressors (cold pressor test, hyperventilation, mental-arithmetic, and word fluency), Lacey et al. (1953) find that different subjects have different patterns of autonomic response which are reproducible over time and are consistent over these different stressors. Thus one subject may respond to the stress by a large increase in heart rate and only a small increase in skin conductance and another may respond with the opposite pattern. To the extent that such findings can be generalized to a clearly fear arousing situation the conclusion is clear that one cannot expect intercorrelations among autonomic change scores to be very substantial. The point to be emphasized here, however, is not that several autonomic measures might not for almost all people increase under anxiety arousing circumstances, but that those measures which show the most or least increase vary from person to person. Such a state of affairs is not necessarily disasterous to one interested in using physiological measures in assessing anxiety. The moral, however, remains clear that for a given individual some physiological measures may be much more sensitive indicators of change in anxiety level than others.

A somewhat similar point of view is espoused by Malmo, Shagass, and Davis (1950) in which they propose the principle of symptom specificity: namely, that psychiatric patients are inclined to respond to stress of all kinds by a particular physiological mechanism that leads to the particular kind of somatic complaint that the patient may have. Thus, Malmo and Shagass (1949b) found that patients with heart complaints showed greater heart rate and heart rate variability under stress than patients without heart complaints. Specificity of muscle potential reaction was demonstrated by Malmo, Smith, and Kohlmeyer (1956) who showed that for the same patient discussion of hostility conflicts was associated with increased forearm muscle tension and discussion of sex conflict was associated with increased leg muscle tension.

There are other studies in which intercorrelations among a number of physiological measures are reported, such as Wenger (1942, 1948) or Gunderson (1953) in which all measures were obtained under resting conditions. If people manifest varying degrees of an autonomic response pattern determined by the amount of anxiety that they "bring into" the resting situation then such a pattern should show up as a cluster of intercorrelated variables. Wenger's (1942) earlier factor analytic work with children did yield a dimension that he called the autonomic factor, which when unbalanced in the sympathetic direction would appear to be similar to the cluster of autonomic measures associated with experimentally aroused anxiety in the previously described studies. However, in Wenger's (1948) study of aviation cadets, operational fatigue patients, and neurotic patients the case for a clear-cut autonomic factor is shaky. The most striking thing about the reported intercorrelations is their extremely low level. Very few correlations are higher than .15. Gunderson (1953), however, reported intercorrelations among his 12 resting state measures on a subsample of 44 paranoid schizophrenics that were both substantial, for this kind of data, and pervasive. There was a tendency for many of the different autonomic measures to correlate between .20 and .45 in a direction consistent with degree of sympathetic arousal.

In summary, intercorrelations among physiological measures obtained under either resting states or under stress tend to be low and frequently insignificant. There are few studies, however, in which a variety of measures are obtained under a clearly fear arousing situation and where the tendency of change scores to correlate with initial level has been partialed out. Improved measurement technique may also make some of the older studies somewhat obsolete. Nevertheless, the best guess on the basis of present findings is that intercorrelations among physiological measures will be found to be low even with the above-mentioned modifications taken into account. Lacey's work suggests, consistent with the findings of low intercorrelations, that an individual responds to stress with a characteristic pattern of responses. This finding may not be entirely inconsistent with the possibility of there being some pattern of response usually associated with fear. For example, Lacey's findings that subjects showed different response patterns to the stress of doing mental arithmetic may result in part from the fact that some subjects were made angry in the situation and some were made anxious, and that those that were made anxious showed a distinctive pattern from those made angry as Funkenstein et al. found. The chances are that this explanation does not account for all the individual response patterns, and it may be that among subjects made anxious there still remain different response patterns. The meaning of these different response patterns, which could be few in number, may be clarified by further knowledge about their correlation with behavioral and perhaps self-report type measures. It is, of course, possible that future factor analytic or pattern analysis studies will suggest the utility of conceptualizing several different kinds of anxiety states.

Behavioral Measures: Experimental and Group Comparisons

The same question is asked here as was asked with respect to physiological measures; is there some pattern of behavioral effects associated with anxiety that can be distinguished from behavioral effects resulting from other arousal states? The researches most relevant to the question are those in the general area of the effects of stress on performance. These researches, unfortunately, do not provide a clear answer to the question because of two major lacks. First, most such studies tend to be limited to one dependent variable for the good reason that it is much more difficult to measure simultaneously a variety of appropriate behavioral responses than physiological responses. Second, few studies attempt to contrast a fear arousal state with other kinds of arousal states. Another general drawback to most behavioral measures for the purposes of assessment, as will be shown in the studies reviewed, is that their relation to the anxiety construct is not a monotonic one; for example, a low score on a certain performance may be associated with a very low or very high state of anxiety. The studies mentioned below, then, can be seen as only suggestive of measures likely to be especially sensitive to the effects of anxiety, and are not intended to represent an extensive coverage of the research on the effects of stress on performance. Summaries of research in this area are provided by Hanfmann (1950), Lazarus, Deese, and Osler (1952), and more recently Easterbrook (1959).

A loose empirical generalization that emerges from studies in this area is that the kinds of tasks most likely to be affected by stress are learning and memory tasks involving novel or relatively poorly learned responses where incorrect competing responses are both numerous and relatively strong; or perceptual tasks in which conditions are imposed that make appropriate discriminations difficult. Thus, failure stress (usually produced by first ego involving, then failing, and then criticizing the subject) has been shown to impair digit span but not vocabulary items (Moldowsky & Moldowsky, 1952); impair recall of incidental learning but not recall of material explicitly instructed to be learned (Aborn, 1953); and impair relearning of a serial list of nonsense syllables (Smith, 1954). Stress imposed by implying that the subject is neurotic or maladjusted on the basis of projective test responses has been found to impair performance on abstract reasoning, ` the Holsopple Concept Formation Test, and mirror tracing (Beier, 1951); and to produce more perseveration of incorrect responses on the Luchins Water Jar Task (Cowen, 1952).

A number of studies in which anxiety is introduced by separating subjects into high and low anxiety groups on the basis of the Taylor MAS (1953) provide evidence not only that the detrimental effect of anxiety becomes greater as the strength and number of incorrect competing responses involved in the task increases, but also, for the levels of anxiety involved, that performance is enhanced for the high anxiety subjects on some tasks when the correct response is very dominant. The incorrect competing responses are usually introduced by increasing the similarity and sometimes also by decreasing the association value of items in a serial learning task, or by both increasing intralist similarity and decreasing similarity between pairs in a paired associate learning task. Lucas (1952), Montague (1953), Farber and Spence (1953), Lazarus, Deese, and Hamilton (1954), Taylor and Chapman (1955), Spence, Farber, and McFann (1956), and Spence, Taylor, and Ketchel (1956), all reported evidence for this relationship. The findings of greater ease of eyeblink conditioning in groups of high anxious as opposed to low anxious subjects (Spence & Farber, 1953; Spence & Taylor, 1951; Taylor, 1951) are also consistent with this general proposition.

Lucas (1952) also studied the effect of experimentally induced failure upon performance as a function of the strength of the incorrect competing responses (manipulated by varying the number of duplications of consonants in a series of consonants being used in an immediate recall task). He found no main effect associated with number of duplications nor any interaction with four degrees of experimentally induced failure. No other studies were found in which anxiety was induced experimentally and its effect upon performance studied where the strength of the incorrect responses was systematically varied within the confines of the same task.

A few studies have made use of real life stress situations that probably meet the need for a really anxiety arousing condition better than the experimental procedures used in the other studies. Beam (1955) obtained measures before doctoral oral examinations and opening night performances in plays as well as at a less stressful period in the subject's life, and found marked impairment in learning a serial list of nonsense syllables, and an increase in palmar sweat and GSR conditioning rate under stress as compared to nonstress. Basowitz et al. (1955) reported a tendency for digit span to be impaired for soldiers undergoing paratroop training as compared to a control group, and Wright (1954) likewise found impairment in digit span in patients confronted with the threat of surgery.

One kind of behavioral measure that would appear promising from an assessment point of view is speech disturbance. Mahl (1956, 1959) has developed a system for reliably scoring speech disturbances of various kinds and has shown certain of these disturbances to be related to variation in anxiety as assessed in psychotherapeutic interviews. Dibner (1956) has employed a similar measure.

In the perceptual area Postman and Bruner (1948) reported impairment in the tachistoscopic perception of three-word sentences under failure stress. Rosenbaum (1953) found greater stimulus generalization under strong shock than weak shock. Smock (1957) reported greater intolerance of ambiguity in a perceptual task under stress than nonstress. Korchin and Basowitz (1954), and Moffitt and Stagner (1956) found increased perceptual closure during paratroop training and experimental threat, respectively.

In studies using group comparisons Angyal (1948) found more impairment in the recognition of patterns of letters under brief exposure conditions in high anxiety patients than other patients. Krugman (1947) and Goldstone (1955) found the threshold for flicker fusion to occur at a lower frequency for anxious than nonanxious subjects.

Eriksen and Wechsler (1955) ingeniously attempted to separate the effects of anxiety (shock induced) on response processes as opposed to sensory discrimination, and concluded that anxiety results in restricted and stereotyped response preferences but does not impair sensory discrimination.

In the studies reviewed so far in this section the effect of stress has been in general to impair performance. There are many studies, however, in which improved performance is associated with stress. Thus, Steisel and Cohen (1951) and Truax and Martin (1957) found improved performance on simple arithmetic problems as a result of failure stress; and Spence (1957) found better recall of words failed on an anagrams task than words successfully completed.

Likewise studies in which groups have been divided on the basis of self-report measures of general anxiety level also indicate that failure stress can lead to improved performance for some subjects. Thus Lucas (1952), Waterhouse and Child (1954), Williams (1955), and Sarason (1956) found that low anxiety subjects tend to improve under stress and high anxiety subjects tend to show impairment under stress.

Thus, to the extent that failure stress arouses anxiety, this construct appears to be associated with both improvement and impairment of performance. These seemingly contradictory findings are in part reconciled in a study by Stennett (1957), who instead of employing just one stress and one nonstress condition attempted to set up four degrees of intensity of motivation. He found that tracking performance improved at first as the rewards for correct performance increased but then showed impairment under the most extreme condition involving a large bonus for high level performance and threat of electric shock if this level was not reached. He also obtained palmar conductance and muscle potential measures on his subjects and found these measures to increase monotonically as a function of increased "motivation." Several authors, consistent with this study and the others previously described, have proposed that adequacy of performance is an inverted U shaped function of some arousal, activation, or emotional state—for example, Woodworth and Schlosberg (1954), Malmo (1957).

Thus there appear to be two rather loose empirical generalizations that can be reached on the basis of the studies reviewed in this section: (*a*) that tasks involving relatively stronger and more numerous competing responses are more subject to the impairing effects of stress, and (*b*) increasing stress results in improved performance up to a point and impairment thereafter. There is no particular evidence in this area to warrant the separation of anxiety as a construct from other more general constructs such as "arousal," "activation," or "drive."

Somewhat differing theoretical formulations have been proposed to account for the empirical generalizations described. Easterbrook (1959) makes a plausible case for the idea that many of the disorganizing effects of emotion can be accounted for on the basis of cue utilization: namely, that increased "drive" or "emotion" leads to a constriction of the perceptual field or decrease in the number of cues that can be attended to. The Iowa theorists, on the other hand (Spence, 1958), employ the concept of drive and its hypothesized multiplicative relationship to habit strength to account for many of the effects of stress on performance; and Child (1954), Child and Waterhouse (1953), and Sarason, Mandler, and Craighill (1952) emphasize the irrelevant competing responses specifically associated with stress on the basis of the past learning.

If anxiety proves to be a distinguishable arousal state, research on its effects on performance would be greatly facilitated if it could be assessed independently, perhaps by physiological measures, from the performance being studied. The utility of this approach is shown in Stennett's study, where it was not necessary to assume that experimental conditions were effective, or to rely upon some paper and pencil measure in determining the presence or magnitude of the motivational or emotional arousal state, but where instead the palmar conductance and muscle potential measures provided more direct evidence of the degree of arousal.

In summary, no studies were discovered in which several objectively measured behavioral characteristics were obtained simultaneously (or almost so) with a variety of physiological measures under conditions likely to be very fear arousing; much less, studies that in addition contrasted different types of arousal states. On the basis of the one and two variable type studies, though, it seems likely that some fairly simple learning, immediate memory, or perceptual tasks could be developed that would be sensitive to changes in anxiety level. It is possible that a few such tasks along with physiological measures could in the future help define more clearly the anxiety response pattern. Although, in general, improved methods of continuous anxiety measurement will probably contribute more to the study of the effects of anxiety on behavior than vice versa.

Behavioral Measures: Intercorrelations

Studies oriented toward assessing the intercorrelations among a number of behavioral manifestations of anxiety are beset by a special problem. Physiological measures can usually be obtained simultaneously but many behavioral effects of anxiety can be assessed only by presenting the subject with a series of tasks to perform. Unknown order effects may well distort the obtained correlations.

There have been several studies of this type in which a number of behavioral measures, selected on the basis of previously reported relationships to anxiety, were intercorrelated. Martin (1958, 1959) in two successive studies using college subjects, found the intercorrelations to be quite low, but a factor analysis still suggested the presence of a dimension that might be labeled anxiety. In the second study some of the measures that had the higher loadings on the anxiety factor were the Taylor MAS, .41; time to learn a complex (five choice) verbal maze, .40; errors in learning of paired associate nonsense syllables with high intralist similarity but low similarity between pairs, .39; tremors on a manual dexterity task, .39; an anxiety check list, .27. A simple verbal maze (two choice) and a paired associate list involving low intralist similarity and high similarity between pairs had zero order loadings on the factor. The loadings with respect to the two kinds of paired associate lists and the two kinds of verbal mazes are consistent with the notion that tasks involving stronger competing responses are more sensitive to the effects of anxiety. A somewhat more prominent factor that also emerged in both studies was interpreted as a motivational factor, that is, a dimension reflecting how hard these college subjects tried on a number of the tasks. Such individual differences in motivation were postulated to be relatively independent of the subjects' anxiety level. A third factor of some generality was identified as intelligence, and yet another factor was entirely defined by self-report measures of anxiety such as the Taylor MAS. Thus performance on a given task such as learning paired associate nonsense syllables with high intralist similarity under mild stress was found not only to be affected by individual differences in anxiety but also by individual differences in motivation, intelligence, and a factor specific to the type of task. Under these circumstances it is easy to see how anxiety variance could frequently be masked by other factors.

Rosenthal (1955), Cattell and Gruen (1955), and Scheier and Cattell (1958) reported several factor analytic studies in which a variety of self-report, behavioral, and, in some cases, physiological measures were obtained. They found a factor, which they label anxiety, emerging in all their studies that is separable from a number of other personality factors after relatively blind rotations to oblique simple structure. The above studies employed substantial *N*s in five different samples of subjects involving USAF pilot trainees, children, and college students. Upon inspection of the factor loadings on the anxiety factor in these various studies as summarized by Cattell and Scheier (1958b) it becomes apparent, however, that the only measures with high loadings and the only measures whose loadings are consistent from study to study are those based on self-report type measures. Few if any behavioral-physiological measures have loadings over .30 and none of those that do are substantiated in any of the other samples. For example, in Rosenthal's study (1955) the three highest loadings on the anxiety factor were Taylor MAS, .85; questionnaire measure of anxious insecurity, .84; and a questionnaire measure of nervous tension, .70. The other four measures with loadings above .30 were also self-report type measures. Rosenthal obtained several physiological measures under various conditions (GSR, heart rate, salivary volume, systolic blood pressure) and none of these were related to this anxiety factor to any degree. Under these circumstances it does not seem reasonable to accept this factor as necessarily assessing the hypothetical anxiety reaction as formulated in this paper.

Cattell and Scheier (1958b) distinguish between the "trait" of anxiety, inferred from factor analysis of a cross section of measures obtained only once on each subject, and the "state" of anxiety inferred from a factor analysis of change scores from one testing time to another. Correlating change scores in this way is referred to as incremental R technique, and Cattell and Scheier (1958a) report in detail the results of such a study. An interesting innovation in this study, too involved to go into in this paper, was the introduction of different "treatment" conditions into a correlational study, so that it was possible to see, for example, how imminence of academic examinations correlated with the other variables. One of the resulting 14 factors was identified as the "state" anxiety factor and appears to represent an arousal state more closely related to the present theoretical view of anxiety than the previously found trait factor. The self-report measures did not dominate the loadings so much, although the two highest loadings were self-report measures involving an anxiety-tension check list, .41, and a questionnaire scale of tension, .40. In addition though, systolic blood pressure had a loading of .30 and palmar conductance of .26. Perhaps inconsistent with this was the positive loading of volume of saliva, .27. The imminence of an academic examination was negatively loaded, −.25, suggesting that just before an examination the usually anxious person becomes less anxious. The authors propose that "a person beset by vague fears and anxieties loses these anxieties for a while when a real danger threatens."

Holtzman and Bitterman (1956) intercorrelated 41 measures obtained on 135 cadets in an Air ROTC unit. These measures included ratings, personality tests, stress tests, perceptual tests, GSR conditioning, and amount of uric acid and glycine in the urine. The intercorrelations among the different kinds of measures were quite low and a factor analysis yielded seven factors which were almost entirely determined by clusters of measures taken from the same test situation.

There are some important limitations to the factor analytic approach to the study of anxiety. For example, there is no convincing logic to the supposition that simple structure, oblique or orthogonal, yields the most psychologically meaningful dimensions; although intuitively it would seem that some kind of oblique solution would be more meaningful for separating out a cluster of physiological-behavioral measures to be idenified as anxiety as opposed to clusters of measures representing other arousal states, since in all likelihood these various arousal states will be correlated. With respect to rotations in factor analytic studies perhaps it would be better if such rotations were not done blindly but with full knowledge of the nature of the measures, and the final rotation considered frankly for what it is, a post hoc hypothesis about the nature of the dimensions revealed. Confirmation of the interpretation of a given factor and further elucidation of the construct validity (Cronbach & Meehl, 1955) of the assessment procedures can then be ascertained by introducing the factor as a variable in experimental research.

Certainly the selection of measures to be intercorrelated affects the definition of the resulting factors. For example, it may be that in the Cattell studies just described, with the exception of the incremental R technique study, that the high intercorrelations among the self-report measures, which almost entirely define the anxiety factor, are due in part to correlated nonanxiety variance. It is also possible that many of the measures used in the factor analytic studies involve characteristic ways of controlling or reducing anxiety rather than more direct manifestations of the anxiety itself. The Holtzman and Bitterman study serves to point up the fact that in an area where correlations between measures obtained from different response systems are going to be low at best, including clusters of highly intercorrelated measures from the same response system or test situation will inevitably result in factors representing these clusters, at least when the common criteria for simple structure are employed. It is possible that a factor analysis done under such conditions might serve to actually hide some real generalities of response, although there is no indication that such was the case in the Holtzman and Bitterman study.

One cannot conclude on the basis of the researches reviewed in this paper, despite many suggestive leads, that any clear-cut pattern of physiological-behavioral responses associated with anxiety arousal, distinguishable from other arousal patterns has been demonstrated. The status of anxiety assessment procedures, both in terms of experimental and correlational findings might be clarified by combining some of the best features of the researches described. First one might attempt to measure simultaneously, or nearly so, an extensive battery of physiological measures and a few selected behavioral measures at a time when the subject is relaxed. This would necessitate a preliminary adaptation-to-the-apparatus session. Then the subjects could be tested again under definitely anxiety arousing circumstances, the more realistic the better. A study of the change score patterns and intercorrelations, after correcting where necessary for correlation with relaxed session levels, should provide evidence for an anxiety pattern if it exists. It would then be further necessary to demonstrate that the pattern of responses was distinguishable from patterns associated with other arousal states such as general activation, anger, or sex; otherwise there is no utility in having a construct of anxiety separate from these others.

When more is known about the physiological-behavioral response pattern associated with anxiety, then self-report scales can be constructed which will predict this response pattern in various situations.

REFERENCES

Aborn, M. The influence of experimentally induced failure on the retention of material acquired through set and incidental learning. *J. exp. Psychol.*, 1953, **45**, 225–231.

Angyal, Alice F. The diagnosis of neurotic traits by means of a new perceptual test. *J. Psychol.*, 1948, **25**, 105–135.

Ax, A. F. The physiological differentiation between fear and anger in humans. *Psychosom. Med.*, 1953, **15**, 433–442.

Barcroft, H., & Konzett, H. On the actions of noradrenaline and isopropyl noradrenaline on the arterial blood pressure, heart rate and muscle blood flow in man. *J. Physiol.*, 1949, **110**, 194–204.

Basowitz, H., Persky, H., Korchin, S. J., & Grinker, R. R. *Anxiety and stress.* New York: McGraw-Hill, 1955.

Beam, J. C. Serial learning and conditioning under real-life stress. *J. abnorm. soc. Psychol.*, 1955, **51**, 543–551.

Beier, E. G. The effect of induced anxiety on flexibility of intellectual functioning. *Psychol. Monogr.*, 1951, **65**(9, Whole No. 326).

Bitterman, M., & Holtzman, W. Conditioning and extinction of the galvanic skin response as a function of anxiety. *J. abnorm. soc. Psychol.*, 1952, **47**, 615–623.

Cattell, R. B., & Gruen, W. The primary personality factors in 11-year-old children by objective tests. *J. Pers.*, 1955, **23**, 460–478.

Cattell, R. B., & Scheier, I. H. Factors in personality change: A discussion of the condition-response incremental design and application to 69 personality response measures and three stimulus conditions. (Advanced Publication No. 9) Urbana, Ill.: Laboratory of Personality Assessment and Group Behavior, University of Illinois, 1958. (a)

Cattell, R. B., & Scheier, I. H. The nature of anxiety: A review of thirteen multivariate analyses comprising 814 variables. *Psychol. Rep.*, 1958, **4**, 351–388. (b)

Child, I. L. Personality. *Annu. Rev. Psychol.*, 1954, **5**, 149–170.

Child, I. L., & Waterhouse, I. K. Frustration and the quality of performance: II. A theoretical statement. *Psychol. Rev.*, 1953, **60**, 127–139.

Clemens, T. L. Autonomic nervous system responses related to the Funkenstein test: I. To epinephrine. *Psychosom. Med.*, 1957, **19**, 267–274.

Cowen, E. L. The influence of varying degrees of psychological stress on problem-solving rigidity. *J. abnorm. soc. Psychol.*, 1952, **47**, 512–519.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. *Psychol. Bull.*, 1955, **52**, 281–302.

Davis, R. C. Response patterns. *Trans. NY Acad Sci., Ser.* II, 1957, **19**, 731–739.

Davis, R. C., & Buchwald, A. M. An exploration of somatic response patterns: Stimulus and sex differences. *J. comp. physiol. Psychol.*, 1957, **50**, 44–52.

Davis, R. C., Buchwald, A. M., & Frank-

MAN, R. W. Autonomic and muscular responses, and their relation to simple stimuli. *Psychol. Monogr.*, 1955, **69**(20, Whole No. 405).

DELARGEY, C., GREENFIELD, A. D. M., McCORRY, R. L., WHELAN, R. F. The effects of intravenous infusion of mixtures of L-adrenaline and L-noradrenaline on the human subject. *Clin. Sci.*, 1950, **9**, 71–78.

DIBNER, A. S. Cue-counting: A measure of anxiety in interviews. *J. consult. Psychol.*, 1956, **20**, 475–478.

DIMASCIO, A., BOYD, R. W., & GREENBLATT, M. Physiological correlates of tension and antagonism during psychotherapy. *Psychosom. Med.*, 1957, **19**, 99–104.

EASTERBROOK, J. A. The effect of emotion on cue utilization and the organization of behavior. *Psychol. Rev.*, 1959, **66**, 183–201.

ERIKSEN, C. W., & WECHSLER, H. Some effects of experimentally induced anxiety upon discrimination behavior. *J. abnorm. soc. Psychol.*, 1955, **51**, 458–463.

FARBER, I. E., & SPENCE, K. W. Complex learning and conditioning as a function of anxiety. *J. exp. Psychol.*, 1953, **45**, 120–125.

FUNKENSTEIN, D. H., GREENBLATT, M., & SOLOMON, H. C. A test which predicts the clinical effects of electric shock treatment on schizophrenic patients. *Amer. J. Psychiat.*, 1950, **106**, 889–901.

FUNKENSTEIN, D. H., GREENBLATT, M., & SOLOMON, H. C. Autonomic changes paralleling psychologic changes in mentally ill patients. *J. nerv. ment. Dis.*, 1951, **114**, 1–18.

FUNKENSTEIN, D. H., GREENBLATT, M., & SOLOMON, H. C. Nor-epinephrine like and epinephrine like substances in psychotic and psychoneurotic patients. *Amer. J. Psychiat.*, 1952, **108**, 652–662.

FUNKENSTEIN, D. H., KING, S. H., & DROLETTE, M. E. *Mastery of stress.* Cambridge: Harvard Univer. Press, 1957.

GOLDENBERG, M. Adrenal medullary function. *Amer. J. Med.*, 1951, **10**, 622–641.

GOLDENBERG, M., PINES, K. L., BALDWIN, E. DE F., GREENE, D. G., & ROH, C. E. The hemodynamic response of man to norepinephrine and epinephrine and its relation to the problems of hypertension. *Amer. J. Med.*, 1948, **5**, 792–799.

GOLDSTONE, S. Flicker fusion measurements and anxiety level. *J. exp. Psychol.*, 1955, **49**, 200–202.

GUNDERSON, E. K. Autonomic balance in schizophrenia. Unpublished doctoral dissertation, University of California, Los Angeles, 1953.

HANFMANN, E. Psychological approaches to the study of anxiety. In P. H. Hoch & J. Zubin (Eds.), *Anxiety.* New York: Grune & Stratton, 1950. Pp. 51–69.

HICKHAM, J. B., CARGILL, W. H., & GOLDEN, G. Cardiovascular reactions to emotional stimuli: Effect on cardiac output, A-V oxygen difference, arterial pressure and peripheral resistance. *J. clin. Invest.*, 1948, **27**, 290–298.

HOLTZMAN, W. H., & BITTERMAN, M. E. A factorial study of adjustment to stress. *J. abnorm. soc. Psychol.*, 1956, **52**, 179–185.

JURKO, M., JOST, H., & HILL, T. S. Pathology of the energy system: An experimental-clinical study of physiological adaptive capacities in a non-patient, a psychoneurotic, and an early paranoid schizophrenic group. *J. Psychol.*, 1952, **33**, 183–198.

KORCHIN, S. J., & BASOWITZ, H. Perceptual adequacy in a life stress. *J. Psychol.*, 1954, **38**, 495–502.

KRUGMAN, M. Flicker fusion frequency as a function of anxiety reaction: An exploratory study. *Psychosom. Med.*, 1947, **9**, 269–272.

LACEY, J. I. Individual differences in somatic response patterns. *J. comp. physiol. Psychol.*, 1950, **43**, 338–350.

LACEY, J. I. The evaluation of autonomic responses: Toward a general solution. *Ann. NY Acad. Sci.*, 1956, **67**, 123–164.

LACEY, J. I., BATEMAN, DOROTHY E., & VAN LEHN, RUTH. Autonomic response specificity: An experimental study. *Psychosom. Med.*, 1953, **15**, 8–21.

LACEY, J. I., & VAN LEHN, RUTH. Differential emphasis in somatic response to stress: An experimental study. *Psychosom. Med.*, 1952, **14**, 71–81.

LAZARUS, R. S., DEESE, J., & HAMILTON, R. Anxiety and stress in learning: The role of intraserial duplication. *J. exp. Psychol.*, 1954, **47**, 111–114.

LAZAZUS, R. S., DEESE, J., & OSLER, S. F. The effects of psychological stress upon performance. *Psychol. Bull.*, 1952, **49**, 293–317.

LEWINSOHN, P. M. Some individual differences in physiological reactivity to stress. *J. comp. physiol. Psychol.*, 1956, **49**, 271–277.

LUCAS, J. D. The interactive effects of anxiety, failure, and intraserial duplication. *Amer. J. Psychol.*, 1952, **65**, 59–66.

MAHL, G. F. Disturbances and silences in the patient's speech in psychotherapy. *J. abnorm. soc. Psychol.*, 1956, **53**, 1–15.

MAHL, G. F. Measuring the patient's anxiety during interviews from "expressive" aspects of his speech. *Trans. NY Acad. Sci.*, 1959, 21, 249–257.

MALMO, R. B. Experimental studies of mental patients under stress. In M. Reymert (Ed.), *Feelings and emotions.* New York: McGraw-Hill, 1950.

MALMO, R. B. Anxiety and behavioral arousal. *Psychol. Rev.*, 1957, **64**, 276–287.

MALMO, R. B., BOAG, T. J., & SMITH, A. A. Physiological study of personal interaction. *Psychosom. Med.*, 1957, **19**, 105–119.

MALMO, R. B., & SHAGASS, C. Physiologic studies of reaction to stress in anxiety and early schizophrenia. *Psychosom. Med.*, 1949, **11**, 9–24. (a)

MALMO, R. B., & SHAGASS, C. Physiologic study of symptom mechanisms in psychiatric patients under stress. *Psychosom. Med.*, 1949, **11**, 25–29. (b)

MALMO, R. B., SHAGASS, C., BELANGER, D. J., & SMITH, A. A. Motor control in psychiatric patients under experimental stress. *J. abnorm. soc. Psychol.*, 1951, **46**, 539–547.

MALMO, R. B., SHAGASS, C., & DAVIS, J. F. A method for the investigation of somatic response mechanisms in psychoneurosis. *Science*, 1950, **112**, 325–328.

MALMO, R. B., SHAGASS, C., & DAVIS, J. F. Electromyographic studies of muscular tension in psychiatric patients under stress. *J. clin. exp. Psychopathol.*, 1951, **12**, 45–66.

MALMO, R. B., & SMITH, A. A. Forehead tension and motor irregularities in psychoneurotic patients under stress. *J. Pers.*, 1955, **23**, 391–406.

MALMO, R. B., SMITH, A. A., & KOHLMEYER, W. A. Motor manifestiation of conflict in interview: A case study. *J. abnorm. soc. Psychol.*, 1956, **52**, 268–271.

MANDLER, G., & KREMEN, I. Autonomic feedback: A correlational study. *J. Pers.*, 1958, **26**, 388–399.

MARTIN, B. A factor analytic study of anxiety. *J. clin. Psychol.*, 1958, **14**, 133–138.

MARTIN, B. The measurement of anxiety. *J. gen. Psychol.*, 1959, **61**, 189–203.

MILLER, N. F. Central stimulation and other new approaches to motivation and reward. *Amer. Psychologist*, 1958, **13**, 100–108.

MOFFITT, J. W., & STAGNER, R. Perceptual rigidity and closure as functions of anxiety. *J. abnorm. soc. Psychol.*, 1956, **52**, 354–357.

MOLDAWSKY, S., & MOLDAWSKY, PATRICA C. Digit span as an anxiety indicator. *J. consult. Psychol.*, 1952, **16**, 115–118.

MONTAGUE, E. K., The role of anxiety in serial rote learning. *J. exp. Psychol.*, 1953, **45**, 91–96.

MOWRER, O. H. A stimulus-response analysis of anxiety and its role as a reinforcing agent. *Psychol. Rev.*, 1939, **46**, 553–565.

POSTMAN, L., & BRUNER, J. S. Perception under stress. *Psychol. Rev.*, 1948, **55**, 314–324.

ROSENBAUM, G. Stimulus generalization as a function of experimentally induced anxiety. *J. exp. Psychol.*, 1953, **45**, 35–43.

ROSENTHAL, IRENE. A factor analysis of anxiety variables. Unpublished doctoral dissertation, University of Illinois, 1955.

SARASON, I. G. Effects of anxiety, motivational instructions, and failure on serial learning. *J. exp. Psychol.*, 1956, **51**, 253–259.

SARASON, S. B., MANDLER, G., & CRAIGHILL, P. G. The effect of differential instructions on anxiety and learning. *J. abnorm. soc. Psychol.*, 1952, **47**, 561–565.

SCHACHTER, J. Pain, fear and anger in hypertensives and normotensives. *Psychosom. Med.*, 1957, **19**, 17–29.

SCHEIER, I. H., & CATTELL, R. B. Confirmation of objective test factors and assessment of their relation to questionnaire factors: A factor analysis of 113 rating, questionnaire, and objective test measurements of personality. *J. ment. Sci.*, 1958, **104**, 608–624.

SCHIFF, E., DOUGAN, C., & WELCH, L. The conditioned PGR and the EEG as indicators of anxiety. *J. abnorm. soc. Psychol.*, 1949, **44**, 549–552.

SELYE, H. *Stress.* Montreal: Acta, 1950.

SHAGASS, C. The sedation threshold: A method for estimating tension in psychiatric patients. *EEG clin. Neurophysiol.*, 1954, **6**, 221–233.

SHAGASS, C. Anxiety, depression, and the photically driven electroencephalogram. *AMA Arch. Neurol. Psychiat.*, 1955, **74**, 3–10. (a)

SHAGASS, C. Differentiation between anxiety and depression by the photically activated electroencephalogram. *Amer. J. Psychiat.*, 1955, **112**, 41–46. (b)

SHAGASS, C., & NAIMAN, J. The sedation threshold, manifest anxiety, and some aspects of ego function. *AMA Arch. Neurol. Psychiat.*, 1955, **74**, 397–406.

SHERMAN, M., & JOST, H. Frustration reactions of normal and neurotic persons. *J. Psychol.*, 1942, **13**, 3–19.

SMITH, J. G. Influence of failure, expressed hostility, and stimulus characteristics on verbal learning and recognition. *J. Pers.*, 1954, **22**, 475–493.

SMOCK, C. D. Recall of interrupted and noninterrupted tasks as a function of experimentally induced anxiety and motivational relevance of the task stimuli. *J. Pers.*, 1957, **25**, 589–599.

Spence, D. P. Success, failure, and recognition threshold. *J. Pers.*, 1957, **25**, 712–720.

Spence, K. W. A theory of emotionally based drive (*D*) and its relation to performance in simple learning situations. *Amer. Psychologist*, 1958, **13**, 131–141.

Spence, K. W., & Farber, I. E. Conditioning and extinction as a function of anxiety. *J. exp. Psychol.*, 1953, **45**, 116–119.

Spence, K. W., Farber, I. E., & McFann, H. H. The relation of anxiety (drive) level to performance in competitional and noncompetitional paired-associates learning. *J. exp. Psychol.*, 1956, **52**, 296–305.

Spence, K. W., & Taylor, Janet A. Anxiety and strength of the US as determiners of the amount of eyelid conditioning. *J. exp. Psychol.*, 1951, **42**, 183–188.

Spence, K. W., Taylor, J., & Ketchel, Rhoda. Anxiety (drive) level and degree of competition in paired-associates learning. *J. exp. Psychol.*, 1956, **52**, 306–310.

Steisel, I. M., & Cohen, B. D. The effects of two degrees of failure on level of aspiration and performance. *J. abnorm. soc. Psychol.*, 1951, **46**, 79–82.

Stennett, R. G. The relationship of alpha amplitude to the level of palmar conductance. *EEG clin. Neurophysiol.*, 1957, **9**, 131–138.

Swan, H. J. C. Noradrenaline, adrenaline, and the human circulation. *Brit. med. J.*, 1952, **1**, 1003–1006.

Taylor, Janet A. The relationship of anxiety to the conditioned eyelid response. *J. exp. Psychol.*, 1951, **41**, 81–92.

Taylor, Janet A., A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, **48**, 285–290.

Taylor, Janet A., & Chapman, Jean P. Anxiety and the learning of paired-associates. *Amer. Psychol.*, 1955, **68**, 671.

Terry, R. A. Autonomic balance and temperament. *J. comp. physiol. Psychol.*, 1953, **46**, 454–460.

Truax, C. B., & Martin, B. The immediate and delayed effect of failure as a function of task complexity and personalization of failure. *J. abnorm. soc. Psychol.*, 1957, **55**, 16–20.

Ulett, G. A., Gleser, G., Winokur, G., & Lawler, A. The EEG and reaction to photic stimulation as an index of anxiety-proneness. *EEG clin. Neurophysiol.*, 1953, 23–32.

Waterhouse, I. K., & Child, I. L. Frustration and the quality of performance: III. An experimental study. *J. Pers.*, 1954, **21**, 298–311.

Welsh, L., & Kubis, J. F. Conditioned PGR (psychogalvanic response) in states of pathological anxiety. *J. nerv. ment. Dis.*, 1947, **105**, 372–381.

Wenger, M. A. A study of physiological factors: The autonomic nervous system and the skeletal musculature. *Hum. Biol.*, 1942, **14**, 69–84.

Wenger, M. A. Studies of autonomic balance in Army Air Forces personnel. *Comp. psychol. Monogr.*, 1948, No. 101.

Wenger, M. A. Pattern analyses of autonomic variables during rest. *Psychosom. Med.*, 1957, **19**, 240–244.

Weybrew, B. B. Patterns of reaction to stress as revealed by a factor analysis of autonomic-change measures and behavioral observation. *J. gen. Psychol.*, 1959, **60**, 253–264.

White, B. V., & Gildea, R. F. "Cold pressor test" in tension and anxiety: A cardiochronographic study. *Arch. Neurol. Psychiat., Chicago*, 1937, **38**, 964–984.

Williams, J. E. Mode of failure, interference tendencies, and achievement imagery. *J. abnorm. soc. Psychol.*, 1955, **51**, 573–580.

Wishner, J. Neurosis and tension: An exploratory study of the relation of physiological and Rorschach measures. *J. abnorm. soc. Psychol.*, 1953, **2**, 253–260

Woodworth, R. S., & Schlosberg, H. *Experimental psychology.* New York: Holt, 1954.

Wright, M. W. A study of anxiety in a general hospital setting. *Canad. J. Psychol.*, 1954, **8**, 195–203.

(Received February 10, 1960)

America's Psychologists

A Survey of a Growing Profession

By Kenneth E. Clark

University of Minnesota

A report of a study of American psychologists, sponsored by the APA Policy and Planning Board and supported by the National Science Foundation, which provides a clearer view of psychology in the mid-twentieth century by describing the people who are active in the field, and the nature of their activities. Some have been outstanding in research productivity. What are they like? How do they differ from their less productive colleagues? Are there major differences among psychologists in, say, experimental psychology and those in, say, industrial psychology? To answer such questions, Dr. Clark and his collaborators have studied the undergraduate education, family backgrounds, types of jobs held, and attitudes and values of different groups of psychologists. How many persons in the United States are engaged in predominantly psychological work? Are recent recipients of the Ph.D. similar to or different from those who received the degree 10 or 20 years ago? Where are psychologists employed? What do they read? These are samples of the questions that are discussed and on which substantial amounts of factual data are given in the pages of this report.

Price, $1.00

★

American Psychological Association
1333 Sixteenth St., N.W.
Washington 6, D.C.

Vol. 58, No. 4 July 1961

Psychological Bulletin

THE SECOND FACET OF FORGETTING: A REVIEW OF WARM-UP DECREMENT

JACK A. ADAMS[1]
University of Illinois

The interference theory of forgetting assumes that the extraexperimental occurrences of S-R sequences, either before original learning of goal responses[2] or interpolated between original learning and recall, will induce a decrement at recall if their stimuli are the same or similar to those of the criterion task and responses are antagonistic. Thus, the laws of forgetting reduce to the laws of proactive and retroactive inhibition (Briggs, 1957; Bugelski & Cadwallader, 1956; Osgood, 1949), with experimental extinction as the process whereby responses are weakened in interference paradigms (Adams, 1952a; Briggs, 1954; Underwood, 1948a, 1948b; Underwood & Postman, 1960). More recently, Underwood (1957) has shown the potency of proactive inhibition on the recall of verbal responses by demonstrating that the prior learning of verbal materials has led us to greatly overestimate the amount forgotten. Underwood's 1957 study, combined with recent research by Underwood and Postman (1960) showing effects on verbal recall from expected sources of verbal interference outside the laboratory, have materially strengthened the interference theory of forgetting. Additional evidence for the interference theory has been by Steinberg and Summerfield (1957) and Summerfield and Steinberg (1957, 1959) who used nitrous oxide in the control of learned associations during interpolated rest. Osgood (1953, pp. 593–597) presents a good review of research in support of the interference theory where various techniques are used to control activities of the organism during the retention interval as a means of reducing opportunities for learning competing responses. Other explanations of normal forgetting might eventually be shown to have validity also, but the preponderance of contemporary evidence lies in support of the interference theory and it will be used without further qualifications throughout this paper as the mechanism by which goal responses are directly influenced and weakened during a retention interval.

The purpose of this paper is to review evidence for the view that warm-up decrement (WU) is a second portion of the retention loss, arising from conditions other than direct interference with goal responses. Irion

[1] Several psychologists read a draft copy of this paper and improved it with their thoughtful commentary. Acknowledgement is due to A. M. Barch, R. C. Davis, M. R. Denny, J. M. Digman, C. P. Duncan, J. C. Jahnke, and B. J. Underwood.

[2] The term "goal response" is used throughout this paper as synonymous with "test response" or "criterial response." It is a feature of overt behavior which the experimenter records and uses as the dependent variable.

(1948) points out that the very special circumstances of stimulus and response similarity required for interference, along with the amount of interfering activity required to produce decrement in the originally learned responses, make it unlikely that the fortuitous experiences of everyday life outside the laboratory could induce significant amounts of one-factor forgetting. While the work of Underwood and Postman (1960) suggests that casual interference is a factor to be reckoned with, there is the strong prevailing sentiment in experimental psychology, supported by research evidence, that hypothesizes WU as a second part of forgetting independent of direct interference with the goal responses. As we shall see, the support for this two-factor view is not as secure as it might be.

Historical Background and Definitions

The first systematic observations on WU arose from interest in fatigue and the characteristics of performance curves under conditions of protracted work, and they appear to have been made in the latter part of the 19th century by Kraepelin and his students (Arai, 1912). Studying a variety of tasks, these researchers observed that the initial segment of a performance curve was typified by a rapid rise in efficiency, followed by a much slower rate of increase or a decline when fatigue effects were present. They identified this initial rapid rise as WU, although in some cases it could have been considered a practice effect or simple reacquisition following one-factor forgetting. Interestingly, these early investigators made an observation which enters the thinking of many later workers: that a rest period contains the simultaneous and opposing processes of beneficial recovery from decremental work effects, and loss of advantageous factors whose reinstatement occurs during the warming-up period.

Mosso (1906) reported anecdotal accounts on the need for poets and writers to warm-up before a period of productive work could begin. Wells (1908) observed the rapid increase in initial postrest performance on a tapping test which, by this time, generally had become identified as WU. Thorndike (1914) in a chapter "Mental Work and Fatigue" gives a more careful definition than previous investigators:

> The best definition of "warming-up" as an an objective act is that part of an increase of efficiency during the first 20 minutes (or some other assigned early portion) of a work period, which is abolished by a moderate rest, say of 60 minutes (p. 66).

One other quotation from Thorndike is particularly significant:

> It should also be noted that intellectual warming-up in the popular sense refers rather to fore-exercise of *other functions*, in order to get materials and motives with which and by which the given function is to work, than to an intrinsic alteration of it (pp. 67–68).

Thorndike's definition of WU as a rapid increase of efficiency during the initial postrest period is consistent with that of earlier writers. Thorndike, in these quotations, makes the influential observation that the WU segment is something other than strengthening of goal responses with practice, and he clearly makes this point in the second quotation (pp. 67–68) when he identifies intellectual WU as the fore-exercise of other functions. It is this identification of WU with factors supporting goal responses which stands as the foundation of the two-factor theory of forgetting, and apparently Thorndike was the first to make it.

Thorndike's observations stand as

the most important historical predecessors of contemporary views, but other investigators made observations on warm-up too, with an occasional experiment. Watson (1919, pp. 354–355) assumed that WU appeared only for heavy muscular work and that the warming-up period was a time of increased glandular action. Robinson and Heron (1924) defined WU as "a rise in efficiency which is steeper and more temporary than the rise which can be seen, let us say, in successive daily performances" (p. 81). Robinson (1934) essentially repeated his 1924 views. Snoddy (1935) presented the first data from a relatively large group of subjects which showed WU following rest. He employed a mirror-tracing instrument as his experimental device. Bell (1942) performed an experiment on the Rotary Pursuit Test (Melton, 1947) on the effects of varying amounts of rest interpolated early and late in practice. Warm-up decrement, as measured by the difference between the first and second postrest trial, was found to first increase and then decrease with amounts of interpolated rest ranging from 1 minute to 30 hours. This trend applied to both early and late in practice.

Post-World War II research displayed an accelerated interest in WU and produced more careful definitions, hypotheses concerning its underlying nature, and specific experimental tests. The modern investigators generally followed the leads of their predecessors. Ammons (1947a), in a miniature system of variables determining rotary pursuit performance, measured WU as the difference between the score on the first postrest trial and a point on the performance curve estimated as the level that would have occurred had there been no need for warming-up. Irion (1948, p. 338) defines WU on the response side in terms of the greater slope of the initial segment of the postrest curve relative to the slope of the original learning curve at a corresponding level of initial proficiency. The response definitions by Ammons and by Irion amount to about the same thing and, along with their theoretical views to be discussed subsequently, have been the mainstay of most workers in the area since the war. A significant feature of these definitions is that they do not imply an actual decrement from the last prerest trial to the first postrest trial and, in this sense, the common reference to "decrement" is a misnomer. Consistent with most early observations on WU, the definitions involve an expression of the sharp initial rise in a postrest performance curve and are independent of whether there is an overall gain or a loss over rest. It is a decrement only in the sense that initial postrest performance is below an expected level because of WU, and this expected level is not always below the level on the final postrest trial. This is the interaction of work and WU effects over rest which drew the attention of Kraepelin and his associates (Arai, 1912). Figure 1 illustrates WU and its appearance under conditions of massed and distributed practice (from Adams, 1952b). The Rotary Pursuit Test was used and 5 days of practice were administered, with 36 ten-second trials given each day. Massed practice was 6 minutes of continuous practice, and distributed practice had a 40-second intertrial rest interval. Eighteen subjects were in the massed group and 21 in the distributed group. These data are a good example of WU manifestations, although a subsequent section

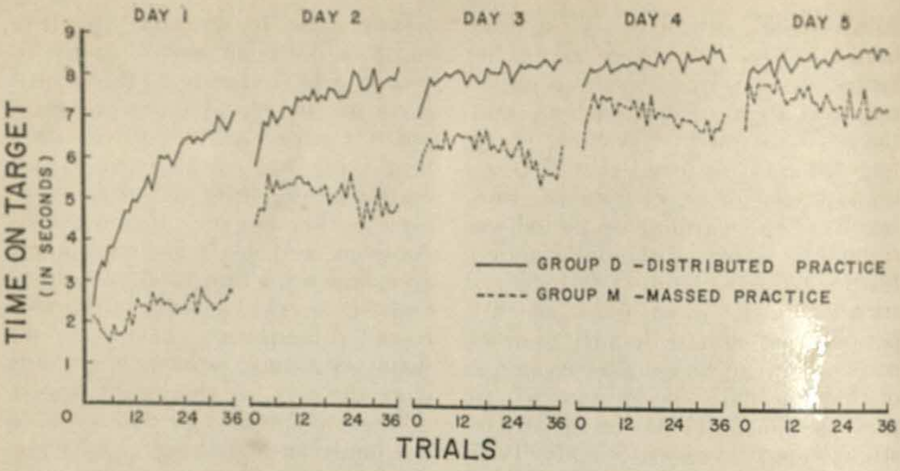


FIG. 1. Illustrations of warm-up decrement under conditions of massed and distributed practice on the Rotary Pursuit Test. (From Adams, 1952b)

will point out that motor WU has a different status at this time than verbal WU. The massed group shows several instances of reminiscence from the final prerest trial to the first postrest trial, but the steep, initial rise in each postrest segment is taken to be WU resulting from a decremental process opposing the gain over rest. Adams measured WU as the difference between the first postrest trial and the score on the trial at the peak of the rise before the decremental segment begins. For the distributed group, however, the method of WU measurement can be the same or it can be measured as a decrement from the last prerest trial to the first postrest trial because reminiscence is absent (Barch, 1954; Reynolds & Adams, 1954). Being able to measure it as an actual decrement is somewhat more precise because it does not involve judgments of the termination point of the WU segment. Digman (1959) replicated Adams' study in most of its aspects and obtained the same trends.

Explanatory Hypotheses

Set

WU must be defined in terms of operations independent of those for a one-factor forgetting interpretation which, for the interference hypothesis of forgetting, would be in terms of responses conflicting with goal responses and causing extinction of them. Considering WU as a performance level below that expected at the beginning of a postrest practice session, it is just as meaningful to regard it as a simple one-factor forgetting loss for the goal response, with WU being a completely superfluous notion. With the exception of Doré and Hilgard (1938) and Hilgard and Smith (1942), pre-World War II investigators demonstrated a lack of methodological caution by simply assuming WU as a phenomenon separate from one-factor forgetting. The Iowa studies of psychomotor interference in the postwar era (Lewis & McAllister, 1950; Lewis, McAllister, & Adams, 1951; Lewis, Shephard, & Adams, 1949; Lewis, Smith, & Mc-

Allister, 1952; Shephard, 1950; Shephard & Lewis, 1950), exhibited a similar conservatism by suggesting an interpretation of WU consistent with the one-factor interference theory of forgetting. They held that the learning of responses in a laboratory task involves the extinction of conflicting responses either from prior tasks learned in the laboratory or from extralaboratory tasks. When a rest period is introduced the extinguished responses spontaneously recover some of their strength and, when postrest practice is resumed, the increased strength of these responses results in heightened conflict with the goal responses and WU occurs. As postrest practice continues the conflicting responses are once again extinguished and WU dissipates. While these one-factor views are parsimonious, and are therefore desirable, the parsimony may be unwarranted. The "other functions" which Thorndike (1914) identified with WU hypothesizes that a one-factor interpretation is an insufficient explanation of WU, and Thorndike's early view is given a more explicit, and testable, expression in the set hypothesis of WU. In the postwar era Irion (1948) gave the first operationally independent statement of WU in terms of a set state of the subject, and this definition was distinct from a one-factor forgetting definition. The term "set" has a number of meanings in psychology (Gibson, 1941) but Irion provided a sufficiently sound operational definition of set within the WU context to provide testable predictions. Although inhibition hypotheses of WU have been attempted, and will be discussed, the set hypothesis has the most status and has been the framework for most of the systematic research on WU.

If set is to be objectively assessed for its utility in the scientific description of behavior, it must be defined in terms of manipulable environmental events, on the one hand, and objective measures of behavior on the other. Furthermore, its operational definition on the environmental side must be different from those defining other behavioral processes which, for our present purposes, is the differentiation of set and one-factor forgetting variables. The independence of defining operations is critical for testing a two-factor theory of forgetting even though forgetting and set loss exert highly similar effects on dependent response measures. Just as long as one-factor forgetting is defined by the retroactive and proactive inhibition paradigm with experimental extinction as the process, and WU is defined by other operations related to a different process such as set, they both can be retained for the description of behavior because they can be independently manipulated and measured. This would be the justification for a scientifically sound two-factor theory of forgetting. Irion's paper (1948) made the two-factor distinction for verbal learning and consequently has given a basis for the objective assessment of set as a determiner of behavior. The use of set with respect to motor behavior has not been grounded in definitions as clear as those for verbal behavior, but this will be discussed later.

Irion's conception of set has much in common with those of Bell (1942) and Ammons (1947a) for motor learning where set is considered to be an aggregate of postural and attentive adjustments which are positively related to performance of the goal response. Complex performance, such as the learning of a verbal list, involves more than the external goal stimuli which the experimenter has objectively defined and controls, and to which the subject links the goal response measured by the experimenter. In addition, various secondary responses are learned, such as the orientation patterns for visual receptors, proper postural attitudes, and muscular tensions. These responses are secondary mainly in the sense of not being directly measured but the efficiency of the goal responding is intimately linked to them. Irion hypothesizes that these secondary responses, or set, are disturbed by the subject's activities between original learning and recall and this loss of set is the underlying cause for the steep slope of the initial segment of the relearning curve which is called WU. The disruption of set could operate to induce the decrement in retention in at least three ways: (*a*) failure of the receptors to adequately receive the goal stimuli, (*b*) mechanical inefficiency for optimum goal responding because the subject does not have the proper posture or muscular tension patterns, and (*c*) change in the internal stimulation which is part of the stimulus complex to which goal responses are conditioned (Guthrie, 1952). This third possible cause of WU could be a function of one or both of the first two because if the secondary responses are disturbed, then their patterns of response-produced stimulation change and the performance level of the goal responses conditioned to these internal cues is reduced.

While set is disrupted by activities during the rest interval, and thus is a kind of interference theory, the hypothesis is distinguished from the one-factor interference theory of forgetting by emphasizing the role of nongoal, secondary responses and, importantly, by specifying that these secondary responses are a function of operations different from those defining the strength of goal responses. The interference theory is concerned with practice variables which strengthen goal S-R sequences and increase their resistance to forgetting, and interfering S-R sequences which weaken goal S-R sequences by experimental extinction. Set, on the other hand, is strengthened by performance of S-R sequences that are neutral with respect to S-R goal sequences and which overcome WU by restrengthening secondary responses—not the strength of goal S-R sequences. While it is true that practice of goal responses appears to strengthen set, as the elimination of WU in relearning testifies, this is only because the goal responses are enmeshed in a matrix of secondary responses and their practice is concurrently accompanied by the practice of secondary responses. However, the task elements which define the learning problem for secondary responses can be embodied in a separate neutral task and can be used to strengthen set independently of goal response practicing. The weakening of set during the retention interval is also presumed to be by interfering activities neutral to goal responses, but their characteristics are unspecified at this time. It might be presumed, for example, that general body movements would disrupt the particular postures and muscular tension patterns acquired in the criterion task.

Of course there may be nothing to the set hypothesis because all retention loss could be one-factor forgetting in terms of direct effects on goal responses. But even given the general terms within which the hypothesis is stated, the scientific criteria are broadly met for verifying that a portion of the retention loss can be ascribed to something other than one-factor forgetting, and it should be possible to find neutral tasks whose performance in a retention interval would reinstate set and abolish WU but would not yield habit strength increments for goal responses. Furthermore, if set is a determiner of performance as Irion says, practice on a neutral task should enhance performance on a criterion task before original learning by strengthening advantageous secondary responses. Effects on recall and original learning will be treated separately in the sections that follow.

Verbal Behavior

Recall. Taking cue from Ward's experiment (Ward, 1937) where the subject's association of colors during rest benefited verbal recall, Irion (1949b) tested the set hypothesis by using one trial of a neutral color-naming task as a warming-up activity just before the recall of paired adjectives after 24 hours. The subject was not required to memorize colors but only to name them as they appeared in the window of a memory drum. Color-naming, then, did not in any way involve practice of goal responses but did serve to reorient the subject to the rhythm of responding and direct his visual attending, posture, and physical adjustments in a manner very similar to that required in the criterion task and should function to restore the subject's set to respond. Irion found that a rest control group which had conventional recall after the 24-hour interval displayed a significant performance loss, but the color-naming group on the first recall trial was significantly superior to the rest control group, and not different from a no-rest control group. This is in good accord with the set hypothesis. In re-establishing performance in the relearning trials, Irion found the one trial of color-naming essentially equivalent to one trial of practice on the criterion task. A related study reported in the same paper demonstrated that first trial recall was a decreasing function of the length of the rest interval up to 24 hours and that the slopes of the relearning curves were a function of the length of the rest interval. Irion interpreted this experiment as being in accord with the set hypothesis and the definition of WU in terms of the slope of the postrest performance curve. Since his color-naming experiment demonstrated that retention loss occurring after 24 hours could be eliminated by one trial of warming-up activity, it seems safe to assume that this decreasing recall function shows increasing WU and loss of set over interpolated time.

Irion and Wham (1951) tested an implication of the set hypothesis that WU should be a decreasing function of the amount of set-reinstating activity. The criterion task was serial rote learning of nonsense syllables and the warming-up activity was recitation of three-place numbers. The retention interval was 35 minutes. Warming-up had a significant effect on the first recall trial, with performance level being a positive function of the number-naming trials. And, rate of increase of the initial WU segment of the relearning curves tended to be inversely related to the amount of warming-up. This study extends Irion's earlier work and represents good support for the set version of WU.

One interpretation that could be

independent variables are manipulating WU through changes in set. Irion (1949a) performed similar research with the Rotary Pursuit Test and related WU to the amount of prerest practice and the duration of rest. His paper reflects the same methodological problem.

Efforts to locate a neutral task which would influence WU in the recall of motor performance, as well as the level of original learning, have met with failure. The most thorough experimental attempt was by Ammons (1951) using the Rotary Pursuit Test. His subjects were administered initial practice, rest, and a postrest practice session. Set-reinstating activities were either watching the disk, blindfolded manual performance of the rotary motion by holding a small rivet set in the rotor plate, or imaginary practice where the subject was merely to think about practicing. These activities were administered either before initial practice, before the postrest practice session, or before both practice periods. No effects were observed from any of the experimental treatments. Walker, DeSoto, and Shelly (1957) performed a bilateral transfer experiment on the Rotary Pursuit Test. Original practice was with one hand and, following rest, practice was resumed with the other hand. One of the experimental conditions was to have one trial of practice just before the postrest session with the prerest practice hand to see if it could have a warming-up effect on the transfer hand. WU was found in performance on the transfer hand but it was unaffected by the warming-up procedure and they concluded that WU must be quite specific to an effector. Hamilton and Mola (1953) used a finger maze and evaluated the effect of practice on five different mazes on performance in a criterion maze. They used an experimental design similar to Hartley (1948) and Thune (1950) and gave practice on the five warming-up mazes either 24 hours or immediately preceding the test maze. Positive transfer to the criterion maze was found but it was about the same for both warm-up groups and the authors concluded that practice on the five mazes exerted only a general practice effect and had no warming-up properties. A small encouraging sign to counter this negative evidence is found in a study by Adams (1955) on sources of work inhibition in complex motor performance. The Rotary Pursuit Test was used and during an intersession period one group was required to observe a partner's performance and press a button every time he judged him to be on target. This activity produced work inhibition but it also tended to result in less WU than found for control groups. The reduced WU was a secondary finding in a study on another topic but it is a lead on a likely set-reinstating activity for motor performance.

Negative Evidence

Ordinarily the amount of evidence which has been cited in support of the set hypothesis for verbal learning would be sufficient to give a good measure of security to the two-factor theory in psychology, but unfortunately there is a disconcerting number of negative findings. In a careful effort to replicate Irion's verbal learning study (1949b), Rockway and Duncan (1952) were unable to reproduce Irion's results and show an effect of color-naming on recall. Similarly, Withey, Buxton, and Elkin (1949) and Hovland and Kurtz (1951) failed to show an influence of color-naming on verbal recall. Underwood (1952), studying the serial learning of nonsense syllables, was unable to demonstrate that the warming-up activity of number-naming prior to recall produced an effect on WU after 24 hours of rest. Underwood said that this did not necesarily contradict previous findings by Irion and others because earlier studies did not use the same subjects in several experimental conditions as he had done. Dinner and Duncan (1959) hypothesized that the unreliability of the effects of color-naming on verbal recall might be a function of degree of original learning. Using a low, medium, and high degree of original learning of paired adjectives, they found that color-naming influenced recall only when level of original learning was high. They concluded that Irion's positive use of color-naming (Irion, 1949b) based on a medium degree of original learning should be considered sampling error and discounted. However, this judgment should be regarded with caution because it does not consider the works of Hartley (1948) and Irion and Wham (1951) who all obtained positive effects of warming-up activities on verbal recall when low or intermediate levels of original learning were used. The Dinner and Duncan investigation makes an original contribution in showing an effect of the amount of original learning, but it cannot be reconciled at this time with verbal learning studies which effectively used warming-up activities to enhance recall.

Another disturbing consideration for understanding WU and the set hypothesis is that there are tasks where performance in the initial postrest trials does not show WU. The inverted alphabet printing task has enjoyed moderate popularity for the study of work inhibition and the data never show WU (Archer, 1954; Eysenck, 1956; Kimble, 1949; Wasserman, 1951). Silver (1952) used the inverted alphabet printing task to investigate the interaction of warm-up activity on WU and work inhibition, but since he used performance on the first postrest trial for his comparisons there is no reason to believe that he was manipulating WU or, indeed whether his data even displayed WU. Other investigators using inverted alphabet printing failed to show WU, and it is unlikely that WU as it has been defined in terms of a rapid increase in performance on the initial postrest trials was even present in Silver's data. Bilodeau (1952a, 1952b), investigated work inhibition by manipulating the physical load required to turn a manual crank, and found no WU in his data. Doten (1955), in a study of interference, used a task where the subject was presented the printed names of colors, but where the color of the letters in the name was different than indicated by the name. The word "Red" might be printed in blue, for example. The task of the subject in original learning was to respond by stating the actual color of the lettering, and no WU was indicated. The initial segments of performance curves on each day had an immediate decrease in speed of responding, not the rapid increase which is characteristic of WU segments. Tasks such as these raise serious definitional problems for set, or any other WU hypothesis for that matter. We cannot expect any explanatory hypothesis to enjoy a good degree of success until the tasks in which WU occurs have been established. Ideally the set hypothesis should contain statements relating WU and task characteristics.

Inhibition

The principal advocate of an inhibition hypothesis is Eysenck (1956) who interprets WU as mainly being attributable to the extinction of Hull's conditioned inhibition (Hull, 1943). Eysenck does not completely deny the set hypothesis, but rather considers loss of set a lesser contributor to WU, with the extinction of conditioned inhibition being the primary reason for the trend of the initial postrest segment. It will be recalled that Hull has a two-factor theory of inhibition. One construct, I_R, is an increasing function of the number of responses and amount of physical work, and a decreasing function of the rest interval. Moreover, I_R has drive properties and its dissipation is regarded as drive reduction. Since drive reduction in Hull's system is the basis of reinforcement, an increment of habit strength for the ongoing response is accrued whenever I_R dissipates. Because the subject is resting when I_R is dissipating, it is theorized that a resting response is strengthened which is antagonistic to the goal response. The habit construct for the resting response is the second inhibitory factor, $_SI_R$. These two types of inhibition summate and subtract from the excitatory potential ($_SE_R$) for the goal response to yield effective excitatory potential ($_S\overline{E}_R$) which is the primary determiner of overt performance level. The massed group in Figure 1 can be used to illustrate Eysenck's application of Hull's inhibition theory to WU. In the first session the subject responds continuously and I_R accrues. Then, over rest, I_R dissipates and an increment of $_SI_R$ develops. The failure of performance on the first postrest trial to reminisce to the level of the distributed group is taken as evidence for the presence of $_SI_R$. When the subject begins practice in the second session the goal response is now being reinforced and the nonreinforced resting response undergoes experimental extinction. This period of extinction of the resting response is revealed as the WU segment, according to Eysenck. One immediate prediction from Eysenck's hypothesis is that little WU should be found under well-spaced practice conditions where a negligible amount of I_R is generated on each trial, and thus negligible $_SI_R$. Eysenck tested this deduction using the Rotary Pursuit Test and, in accord with his prediction, found WU under conditions of massed practice but not distributed practice. His findings and conclusions are tenuous however, because of the instances in the experimental literature showing WU under conditions of widely distributed practice on the Rotary Pursuit Test. Figure 1 is a good example (Adams, 1952b). Other examples are Ammons (1950), Denny, Frisbey, and Weaver (1955), Digman (1959), Kimble and Shatel (1952), and Jahnke and Duncan (1956). There is no immediate explanation for Eysenck's unusual finding, but WU under conditions of well-distributed practice is a commonplace finding and suggests that Eysenck's hypothesis cannot be taken seriously.

Adams (1952b) entertained a different inhibition hypothesis. Observing that much of the evidence for WU came from studies on the Rotary Pursuit Test under conditions of massed practice, he deduced the characteristics of a postrest performance curve from a negatively accelerated growth of reaction potential with trials and an ogival function for the accrual of work inhibition when practice is massed. It was predicted that WU should not appear under conditions of distributed practice. The occurrence of clearcut WU when training was widely distributed (Figure 1) led to rejection of the hypothesis.

At present it must be concluded that no convincing evidence exists in support of an inhibition explanation of WU.

Warm-Up in Animals

The main concern over WU has been with human subjects, but it is noteworthy that the phenomenon also has been observed in animals. The following studies are not meant to represent an exhaustive search of the literature on animal behavior, but rather are intended to show the ubiquity of WU-like effects and that its characteristics are not found only in human response records. Schlosberg (1934, 1936) interprets as WU the failure of occurrence of a well-learned conditioned response in the white rat on the first few trials of a learning session. Ellson (1938), in studying extinction of a bar-pressing habit in the rat, found rate of response slower in the first fifth of the extinction trials than in the second fifth. He interpreted this as WU and explained it in terms of Guthrie's theory which holds that the stimuli to which a response is learned include internal stimuli resulting from posture, movement, etc. Later responses are partly conditioned to the response-produced stimuli of earlier responses and we would expect that later responses in a series would have greater strength because of the presence of the response-produced stimuli to which they are conditioned. In the latter part of extinction the effects of nonreward overcome this trend and the performance level then decreases systematically. Finger (1942) used rats in a straight runway situation and found WU revealed when an extinction series was administered after 24 hours. The second extinction trial actually had performance superior to that of the first extinction trial—a finding contrary to expectation for an extinction series. Finger's finding for extinction is quite similar to Ellson's. Verplanck (1942) reported WU for rats in a simple running task. Like many investigators of human behavior, these animal researchers freely labeled decrements in initial postrest segments of performance records as WU although the decrements could just as well, and more economically, have been explained by the one-factor forgetting hypothesis.

Discussions and Conclusions

Virtually all support for the two-factor theory of forgetting is embodied in experiments which have demonstrated that WU is reduced or eliminated by repetition of responses that orient the subject to the general task demands (e.g., color-naming) but which do not involve direct practice of goal responses. By themselves these experiments might be sufficient to establish set as a second factor necessary for the explanation of retention loss, but the studies where set-reinstating activities have failed to influence recall in both motor and verbal tasks, and the tasks where no WU whatsoever has been found, leave the second factor in doubt. There is not sufficient evidence to reject the set hypothesis but neither are there grounds for firmly retaining it. Certainly it is the most tenable of all hypotheses advanced, but a great deal of careful research seems required before the set hypothesis, and thus the two-factor theory of forgetting, can be accepted or rejected with confidence.

It is unlikely that a decision ever

will be made about the set hypothesis unless it receives a more thorough testing than it has in the past. The set-reinstating experiments are very broadly derived from the hypothesis and have not been a test of the more explicit implications of set. By viewing the set hypothesis in its more detailed aspects, and in attempting to develop specific experiments and measures to empirically verify these details, it should be possible not only to clarify the status of the set hypothesis but also to determine why, for example, some tasks display WU and others do not. For example, Irion (1948), Ammons (1947a), and Bell (1942) all contend that the acquisition of set can be the learning of beneficial postures and muscular tensions that facilitate the occurrence of goal response sequences. Rest period activities disturb the favorable set and the WU segment of a postrest performance curve represents the reacquisition of these favorable bodily attitudes. If there is anything to this version of the set hypothesis, it would seem fruitful to explore the characteristics of bodily tensions by direct measurement and then relate it to changes in performance of the goal response. Davis and his associates have performed a number of studies (e.g., Davis, 1940, 1956) showing the relationship between the characteristics of overt responding and muscular tensions as revealed by electromyographic measurement techniques. Davis (1956) does not believe that the muscular substrata and the overt goal responses need be conceptualized as fundamentally different. A state of tension in skeletal muscle is the same as any other muscular contraction, i.e., it is a response configuration. Davis (1956) says:

> Muscular tensions would then be themselves responses to stimuli, many being small responses, detectable only with instruments, but with no firm boundary between them and the larger muscular activities associated with movement (p. 2).

Davis' work is suggestive for the set hypothesis because it strongly hints that the pattern of electromyographic measures of muscular tension during the WU segment of the postrest performance curve would have levels and patterns of muscular tension different from final prerest performance, and these levels and patterns will have shifted in the direction associated with poorer performance. Moreover, the reacquisition of prerest values and patterns of muscular tensions should parallel the trend of the WU segment. Furthermore, and importantly, it suggests that neutral set-reinstating activities will produce electromyographic changes signifying that the favorable muscular tensions existing at final prerest performance are being re-established.

There are difficulties in operationally distinguishing between an electromyographically-verified muscle tension version of the set hypothesis and Irion's alternative Guthrian hypothesis that loss of set is disturbance of internal stimuli to which goal responses are partly conditioned. If we have changes in the muscle tension secondary responses and this in turn, results in a lower level for goal responses, we cannot be sure that the lower level is due to quasi-mechanical considerations where muscular tensions underlie useful postures and bodily attitudes, or whether it is due to changes in the population of stimuli to which the goal responses have been conditioned. Despite the potential difficulties of interpreting the primary effects of muscle tension secondary responses, on performance of the goal responses, it would be a fundamental finding to show a systematic covariation of electromyographic measures and WU phenomena. The evanescent quality of set could benefit from a diversity of approaches at this time to provide clues for a reconciliation of inconsistencies among the various experimental findings.

The delineation of set and its role in retention will sharpen our understanding of the retention loss problem and will improve our efforts to predict and control it. Underwood (1957) has shown that our frequent use of the same subjects in several laboratory experiments has led us to greatly overestimate the retention loss for verbal responses because the experimenter was unwittingly contaminating his retention scores with proactive inhibition effects. But even given this downward revision of retention loss, we are still faced with showing the proportion of it attributable to interference with goal responses and the part assignable to loss of set. Irion's experiment (1949b), for example, showed that one trial of color-naming almost completely eliminated the verbal retention loss and therefore all of the loss could be described in terms of change in set. This suggests that if the two-factor theory eventually becomes better established in fact the paradigm of retention studies will have to include groups whose performance of set-reinstating activities will allow a parsing of set and interference components. Interference with goal responses may be a smaller contributor to retention loss than we now surmise. The first research need however, is a more incisive laboratory attack on the validity of set and its underlying nature.

REFERENCES

Adams, J. A. The influence of the time interval after interpolated activity on psychomotor performance. *USAF Hum. Resour. Res. Cent. res. Bull.*, 1952, No. 52-11, (a)

Adams, J. A. Warm-up decrement in performance on the pursuit-rotor. *Amer. J. Psychol.*, 1952, **65**, 404–414. (b)

Adams, J. A. A source of decrement in psychomotor performance. *J. exp. Psychol.*, 1955, **49**, 390–394.

Ammons, R. B. Acquisition of motor skill: I. Quantitative analysis and theoretical formulation. *Psychol. Rev.*, 1947, **54**, 263–281. (a)

Ammons, R. B. Acquisition of motor skill: II. Rotary pursuit performance with continuous practice before and after a single rest. *J. exp. Psychol.*, 1947, **37**, 393–411. (b)

Ammons, R. B. Acquisition of motor skill: III. Effects of initially distributed practice on rotary pursuit performance. *J. exp. Psychol.*, 1950, **40**, 777–787.

Ammons, R. B. Effects of prepractice activities on rotary pursuit performance. *J. exp. Psychol.*, 1951, **41**, 187–191.

Arai, T. Mental fatigue. *Teach. Coll. Contri. Educ.*, 1912, No. 54.

Archer, J. E. Postrest performance in motor learning as a function of prerest degree of distribution of practice. *J. exp. Psychol.*, 1954, **47**, 47–51.

Barch, A. M. Warm-up in massed and distributed pursuit rotor performance. *J. exp. Psychol.*, 1954, **47**, 357–361.

Bell, H. M. Rest pauses in motor learning as related to Snoddy's hypothesis of mental growth. *Psychol. Monogr.*, 1942, **54**(1, Whole No. 243).

Bilodeau, E. A. Decrements and recovery from decrements in a simple work task with variation in force requirements at different stages of practice. *J. exp. Psychol.*, 1952, **44**, 96–100. (a)

Bilodeau, E. A. Massing and spacing phenomena as a function of prolonged and extended practice. *J. exp. Psychol.*, 1952, **44**, 108–113. (b)

Briggs, G. E. Acquisition, extinction, and recovery function in retroactive inhibition. *J. exp. Psychol.*, 1954, **47**, 285–293.

Briggs, G. E. Retroactive inhibition as a function of the degree of original and interpolated learning. *J. exp. Psychol.*, 1957, **53**, 60–67.

Bugelski, B. R., & Cadwallader, T. C. A

reappraisal of the transfer and retroaction surface. *J. exp. Psychol.*, 1956, **52**, 360–366.

Davis, R. C. Set and muscular tension. *Indiana U. Publ., Sci. Ser.*, 1940, No. 10.

Davis, R. C. Electromyographic factors in aircraft control: The relation of muscular tension to performance. *USAF Sch. Aviat. Med. Rep.*, 1956, No. 55–122.

Denny, R. M., Frisbey, N., & Weaver, J., Jr. Rotary pursuit performance under alternate conditions of distributed and massed practice. *J. exp. Psychol.*, 1955, **49**, 48–54.

Digman, J. M. Growth of a motor skill as a function of distribution of practice. *J. exp. Psychol.*, 1959, **57**, 310–316.

Dinner, Judith E., & Duncan, C. P. Warm-up in retention as a function of degree of verbal learning. *J. exp. Psychol.*, 1959, **57**, 257–261.

Doré, L. R., & Hilgard, E. R. Spaced practice as a test of Snoddy's two processes in mental growth. *J. exp. Psychol.*, 1938, **23**, 359–374.

Doten, G. W. The effects of rest periods on interference of a well-established habit. *J. exp. Psychol.*, 1955, **49**, 401–406.

Ellson, D. G. Quantitative studies of the interaction of simple habits: I. Recovery from specific and generalized effects of extinction. *J. exp. Psychol.*, 1938, **23**, 339–358.

Eysenck, H. J. "Warm-up" in pursuit rotor learning as a function of the extinction of conditioned inhibition. *Acta psychol., Amst.*, 1956, **12**, 349–370.

Finger, F. W. Retention and subsequent extinction of a simple running response following varying conditions of reinforcement. *J. exp. Psychol.*, 1942, **31**, 120–133.

Gibson, J. J. A critical review of the concept of set in contemporary experimental psychology. *Psychol. Bull.*, 1941, **38**, 781–817.

Guthrie, E. R. *The psychology of learning.* (Rev. ed.) New York: Harper, 1952.

Hamilton, C. E. The relationship between length of interval separating two learning tasks and performance on the second task. *J. exp. Psychol.*, 1950, **40**, 613–621.

Hamilton, C. E., & Mola, W. R. Warm-up effect in human maze learning. *J. exp. Psychol.*, 1953, **45**, 437–441.

Hartley, T. C. Retention as a function of the temporal position of an interpolated warming-up task. Unpublished MA thesis, University of Illinois, 1948.

Heron, W. T. The warming-up effect in learning nonsense syllables. *J. genet. Psychol.*, 1928, **35**, 219–228.

Hilgard, E. R., & Smith, M. B. Distributed practice in motor learning: Score changes within and between daily sessions. *J. exp. Psychol.*, 1942, **30**, 136–146.

Hovland, C. I., & Kurtz, K. H. Experimental studies in rote-learning theory: IX. Influence of work-decrement factors on verbal learning. *J. exp. Psychol.*, 1951, **42**, 265–272.

Hull, C. L. *Principles of behavior.* New York: Appleton-Century, 1943.

Hunter, I. A. The warming-up effect in recall performance. *Quart. J. exp. Psychol.*, 1955, **7**, 166–175.

Irion, A. L. The relation of "set" to retention. *Psychol. Rev.*, 1948, **55**, 336–341.

Irion, A. L. Reminiscence in pursuit-rotor learning as a function of length of rest and of amount of pre-rest practice. *J. exp. Psychol.*, 1949, **39**, 492–499. (a)

Irion, A. L. Retention and warming-up effects in paired associate learning. *J. exp. Psychol.*, 1949, **39**, 669–675. (b)

Irion, A. L., & Wham, Dorothy S. Recovery from retention loss as a function of amount of pre-recall warming-up. *J. exp. Psychol.*, 1951, **41**, 242–246.

Jahnke, J. C., & Duncan, C. P. Reminiscence and forgetting in motor learning after extended rest intervals. *J. exp. Psychol.*, 1956, **52**, 273–282.

Kimble, G. A. An experimental test of two-factor theory of inhibition. *J. exp. Psychol.*, 1949, **39**, 15–23.

Kimble, G. A., & Shatel, R. B. The relationship between two kinds of inhibition and the amount of practice. *J. exp. Psychol.*, 1952, **44**, 355–359.

Lewis, D., & McAllister, Dorothy E. An investigation of individual susceptibility to interference. *USN Spec. Dev. Cent. tech. Rep.*, 1950, No. 938-1-10.

Lewis, D., McAllister, Dorothy E., & Adams, J. A. Facilitation and interference in performance on the modified Mashburn apparatus: I. The effects of varying the amount of original learning. *J. exp. Psychol.*, 1951, **41**, 247–260.

Lewis, D., Shephard, A. H., & Adams, J. A. Evidences of associative interferences in psychomotor performance. *Science*, 1949, **110**, 271–273.

Lewis, D., Smith, P. N., & McAllister, Dorothy E. Retroactive facilitation and interference in performance on the Modified Two-Hand Coordinator. *J. exp. Psychol.*, 1952, **44**, 44–50.

MELTON, A. W. (Ed.) *Apparatus tests.* (AAF Aviat. Psychol. Program res. Rep. No. 4) Washington, D. C.: United States Government Printing Office, 1947.

MOSSO, A. *Fatigue.* (Trans. by M. Drummond) New York: Putnam, 1906.

OSGOOD, C. E. The similarity paradox in human learning. *Psychol. Rev.*, 1949, **56**, 132–143.

OSGOOD, C. E. *Method and theory in experimental psychology.* New York: Oxford, 1953.

REYNOLDS, B., & ADAMS, J. A. Psychomotor performance as a function of initial level of ability. *Amer. J. Psychol.*, 1954, **67**, 268–277.

ROBINSON, E. S. Work of the integrated organism. In C. Murchison (Ed.), *Handbook of general experimental psychology*, 1934.

ROBINSON, E. S., & HERON, W. T. The warming-up effect. *J. exp. Psychol.*, 1924, **7**, 81–97.

ROCKWAY, M. R., & DUNCAN, C. P. Pre-recall warming-up in verbal retention. *J. exp. Psychol.*, 1952, **43**, 305–312.

SCHLOSBERG, H. Conditioned responses in the white rat. *J. genet. Psychol.*, 1934, **45**, 303–335.

SCHLOSBERG, H. Conditioned responses in the white rat: II. Conditioned responses based upon shock to the foreleg. *J. genet. Psychol.*, 1936, **49**, 107–138.

SHEPHARD, A. H. Losses of skill in performing the standard Mashburn task arising from different levels of learning on the reversed task. *USN Spec. Dev. Cent. tech. Rep.*, 1950, No. 938-1-9.

SHEPHARD, A. H., & LEWIS, D. Prior learning as a factor in shaping performance curves. *USN Spec. Dev. Cent. tech. Rep.*, 1950, No. 938–1–4.

SILVER, R. J. Effect of amount and distribution of warming-up activity on retention in motor learning. *J. exp. Psychol.*, 1952, **44**, 88–95.

SNODDY, G. S. *Evidence for two opposed processes in mental growth.* Lancaster: Science, 1935.

STEINBERG, HANNAH, & SUMMERFIELD, A. Influence of a depressant drug on acquisition in rote learning. *Quart. J. exp. Psychol.*, 1957, **9**, 138–145.

SUMMERFIELD, A., & STEINBERG, HANNAH. Reducing interference in forgetting. *Quart. J. exp. Psychol.*, 1957, **9**, 146–154.

SUMMERFIELD, A., & STEINBERG, HANNAH. Using drugs to alter memory experimentally in man. In P. B. Bradley, P. Deniker, & C. Radouco-Thomas (Eds.), *Neuro-psychopharmacology.* Houston: Elsevier, 1959. Pp. 481–483.

THORNDIKE, E. L. *Educational psychology.* Vol. 3. New York: Teachers Coll., Columbia Univer., 1914.

THUNE, L. E. The effects of different types of preliminary activities on subsequent learning of paired-associate learning. *J. exp. Psychol.*, 1950, **40**, 423–438.

THUNE, L. E. Warm-up effect as a function of level of practice in verbal learning. *J. exp. Psychol.*, 1951, **42**, 250–256.

UNDERWOOD, B. J. Retroactive and proactive inhibition after five and forty-eight hours. *J. exp. Psychol.*, 1948, **38**, 29–38. (a)

UNDERWOOD, B. J. "Spontaneous recovery" of verbal associations. *J. exp. Psychol.*, 1948, **38**, 429–439. (b)

UNDERWOOD, B. J. Studies of distributed practice: VI. The influence of rest-interval activity in serial learning. *J. exp. Psychol.*, 1952, **43**, 329–340.

UNDERWOOD, B. J. Interference and forgetting. *Psychol. Rev.*, 1957, **64**, 49–60.

UNDERWOOD, B. J., & POSTMAN, L. Extra-experimental sources of interference in forgetting. *Psychol. Rev.*, 1960, **67**, 73–95.

VERPLANCK, W. S. The development of discrimination in a simple locomotor habit. *J. exp. Psychol.*, 1942, **31**, 441–464.

WALKER, L. C., DESOTO, C. B., & SHELLY, M. W. Rest and warm-up in bilateral transfer on a pursuit rotor task. *J. exp. Psychol.*, 1957, **53**, 394–404.

WARD, L. B. Reminiscence and rote learning. *Psychol. Monogr.*, 1937, **49**(4, Whole No. 220).

WASSERMAN, H. N. The effect of motivation and amount of pre-rest practice upon inhibitory potential in motor learning. *J. exp. Psychol.*, 1951, **42**, 162–172.

WATSON, J. B. *Psychology from the standpoint of a behaviorist.* Philadelphia: Lippincott, 1919.

WELLS, F. L. Normal performance on the tapping test before and during practice with special reference to fatigue phenomenon. *Amer. J. Psychol.*, 1908, **19**, 437–483.

WHITHEY, S., BUXTON, C. E., & ELKIN, A. Control of rest interval activities in serial verbal learning. *J. exp. Psychol.*, 1949, **39**, 173–176.

(Received July 15, 1960)

Psychological Bulletin
1961, Vol. 58, No. 4, 274–298

ON THE REFORMULATION OF INHIBITION IN HULL'S SYSTEM

ARTHUR R. JENSEN
University of California

Among the least satisfactory elements of Hull's behavior system is his formulation of inhibition. As a result, there have been several attempts in recent years to reformulate Hull's theory with respect to the inhibition variables in the equation for effective reaction potential (${}_S\bar{E}_R$). The present paper critically examines these reformulations in the light of relevant experimental evidence. The conclusions to which this examination leads are that these reformulations have not been an improvement over Hull and that this kind of reformulation itself is a futile approach to the problem of improving Hullian-type learning theory.

In all versions of his theory Hull (1943, 1951, 1952) formulated "effective reaction potential" (${}_S\bar{E}_R$) as being essentially a function of "drive" (D) and "habit strength" (${}_SH_R$), related multiplicatively (i.e., $D \times {}_SH_R$), *minus* "reactive inhibition" (I_R) and "conditioned inhibition" (${}_SI_R$), related additively (i.e., $I_R + {}_SI_R$). Thus:

$${}_S\bar{E}_R = (D \times {}_SH_R) - (I_R + {}_SI_R)$$

Most of the attempts to reformulate Hull's equation have been the result of logical, or at times merely verbal, rather than empirical considerations. For example, Hilgard's (1956, p. 139) criticism is directed at the fact that Hull did not carry out the logical implications of his statement that I_R is a "negative drive state." As such, I_R logically should subtract from D (i.e., $D - I_R$) and, like D, should interact multiplicatively with habit strength (i.e., $I_R \times {}_SH_R$). Hilgard also suggests that, since ${}_SI_R$ is a negative habit, it should interact multiplicatively with I_R. Thus, Hilgard's proposed reformulation of the equation for net reaction potential results in the following:

$${}_S\bar{E}_R = [(D - I_R) \times {}_SH_R] - (I_R \times {}_SI_R)$$

This new formulation seems to be more consistent with some of Hull's own statements about the nature of these intervening variables, but Hilgard avoids trouble by not attempting to relate this formulation to empirical findings.

Similarly, Iwahara (1957) carries Hull's characterization of I_R as a negative drive and ${}_SI_R$ as a negative habit to what may seem the logical conclusion in terms of the internal consistency of Hull's theory—that the relationship between drives and habits is always multiplicative and never additive. Iwahara then goes a step further to regard ${}_SI_R$ as a conditioned or secondary negative drive, with I_R being the primary negative drive. From this it follows that the product of $I_R \times {}_SI_R$ should subtract from positive drive, D, and should also multiply ${}_SH_R$. Symbolically,

$${}_S\bar{E}_R = {}_SH_R \times [D - (I_R \times {}_SI_R)]$$

or, in expanded form,

$${}_S\bar{E}_R = ({}_SH_R \times D) - ({}_SH_R \times I_R \times {}_SI_R)$$

Osgood (1953, p. 379) states that Hull need not have postulated ${}_SI_R$ at all, since it might have been derived from other postulates in the system. If ${}_SI_R$ is nothing other than negative habit strength or the habit of not responding (reinforced by the dissipation of I_R), it would seem logical to subtract $_SI_R$ directly from $_SH_R$. This is the formulation Osgood has proposed (p. 349).

More recently, Jones (1958) has incorporated the foregoing suggestions in his revision of Hull's equation. The Jones version, which combines the properties of the other revisions (except Iwahara's $_SH_R \times {_SI_R}$) and appears identical to Osgood's suggestion, is as follows:

$$_S\overline{E}_R = (D - I_R) \times (_SH_R - {_SI_R})$$

That this formulation is quite radically different from Hull's is even more obvious when Jones mathematically expands the equation, thus:

$$_S\overline{E}_R = (D \times {_SH_R}) - (I_R \times {_SH_R}) - (D \times {_SI_R}) + (I_R \times {_SI_R})$$

Jones' formulation has been subscribed to by Eysenck and his coworkers in their attempt to utilize Hullian postulates in developing a theory of personality (Eysenck, 1957; Kendrick, 1958).

Another revision, rather casually suggested by Woodworth and Schlosberg (1954, p. 668), is that inhibition (I_R or $_SI_R$ or both?) should subtract from "incentive motivation" (Hull's K, a function of the amount of reinforcement). Presumably the total inhibitory potential I_R (the sum of $I_R + {_SI_R}$) subtracts from K, though this point is not clear in the Woodworth and Schlosberg discussion. Their suggestion might be expressed symbolically as follows:

$$_S\overline{E}_R = (K - I_R - {_SI_R}) \times D \times {_SH_R}$$

The most carefully formulated and empirically anchored modifications of Hull's theory have been those of Spence (1956). His changes in the inhibition part of the theory are of a fundamentally different nature than the other revisions. He has more or less wiped the slate clean and started anew by redefining inhibition and the independent variables of which it is a function. Spence's extinctive inhibition (I_n) is not a function of the amount of effort or rate of responding, as is Hull's I_R, but is a function only of the number of nonreinforced responses. There is also an oscillatory inhibition (I_o), which is the same as Hull's concept of oscillation ($_SO_R$). The inhibition due to delay of reward (I_t) is essentially the same as I_n. The basis of this inhibition is assumed to be the competing responses that are established during the delay period or during extinction. The molar concepts of I_t or I_n simply represent the quantitative effects of these competing responses. Spence's inhibition does not interact with other intervening variables but only subtracts from the reaction potential. In this last respect his formulation is essentially no different from Hull's. It might be asked why D, if it is regarded as an energizer of all responses in the organism's repertoire, should *not* interact with inhibition as Spence conceives of it, that is, as consisting of interfering or competing responses. In this respect Spence's theory of extinction is not unlike Guthrie's.

With the exception of Spence, these attempts to reformulate Hull raise a number of crucial questions in common, some of which must be critically examined on the level of theory and methodology and others in terms of empirical evidence. First there are questions of a general theoretical nature which must be considered in relation to any attempt to criticize or reformulate Hull's theory.

1. Is the verbal formulation of

Hull's theory to be taken more seriously than the symbolic and quasi-quantitative formulations, or than the actual empirical relationships which formed the basis for Hull's postulates and which he has held up as examples of the relationships he wished his system to predict?

2. Does the algebraic manipulation of Hull's intervening variables make sense theoretically and psychologically? Are the functions representing their interrelationships "isomorphic" with the rules of simple algebra?

3. Can experiments be designed to determine the exact nature of the intervening variables?

Once one has decided to argue within the Hullian framework a number of questions arise from the attempts at reformulation, the answers to which must depend upon empirical findings.

1. Does ${}_SI_R$ subtract from ${}_SH_R$? Are ${}_SH_R$ and ${}_SI_R$ both basically the same phenomenon, one merely being positive and the other negative in effect, or do they represent basically different processes?

2. Is there any empirical evidence to support the following formulations?

a. The interaction of $D \times {}_SI_R$ (Jones, Osgood)

b. $D - I_R$ (Hilgard, Jones, Osgood)

c. The interaction of ${}_SH_R \times I_R$ (Hilgard, Iwahara, Jones, Osgood)

d. The interaction of ${}_SH_R \times {}_SI_R$ (Iwahara)

e. The interaction of $I_R \times {}_SI_R$, which paradoxically represents an *addition* to reaction potential, the multiplication of two negative quantities making a positive (Hilgard, Iwahara, Jones, Osgood)

f. $K - I_R$ (Woodworth & Schlosberg)

The Limitations of Hull's Theory

In offering his revision, Jones (1958) points out that the inhibition aspect of Hull's formula for reaction potential has been criticized by Koch (1954). Koch's criticisms, however, apply equally to Jones' revision as well as to all the others, with the possible exception of Spence. Koch points out that the intervening variables concerning inhibition in Hull's system, particularly ${}_SI_R$, are not rigorously defined, are not clearly tied to experimental variables, and hence are indeterminate. Because of this, it is impossible to make rigorous experimental tests of Hull's formulations or of the alternative revisions. Cotton (1955) has shown that a literal interpretation of Hull's postulates leads to predictions that differ from the experimental data upon which Hull based the formulation of his postulates in the first place. In short, much of Hull's theory does not even predict the very facts it was expressly devised to predict. This is especially true with regard to the inhibition postulates. None of the revisions of Hull has improved this situation. They have merely rearranged in various ways the same indeterminate variables of Hull's formula for ${}_S\overline{E}_R$.

Hull's revisers have followed him in treating his intervening variables, D, ${}_SH_R$, I_R, ${}_SI_R$, etc., as if they were real, independent quantities whose laws of interaction are isomorphic with the rules of arithmetic and algebra. As we shall see, the manipulation of these hypothetical variables in such fashion can at times lead to absurdity. Hull's intervening variables *are* only intervening variables in the sense which MacCorquodale and Meehl (1948) have assigned to that term, and are defined only in terms of the independent and dependent variables to which they are tied. The danger arises when Hull's revisers mathematically manipulate the intervening variables without regard for the defining experimental variables which are actually all that give any meaning to the intervening variables. Of course, one of the purported virtues of intervening variables is that they can be mathematically manipulated as independent entities. But once the intervening variable has been properly defined, the question arises as to the nature of the mathematical operations that can suitably be applied to it. It is highly doubtful if the exclusive use of linear algebra by Hull and his revisers is at all suitable. It should be noted that in Hull's own statements (1943) the relationship between experimental variables and intervening variables is usually anything but linear. If the exact form of the functional relationship is not known, performing linear algebraic operations on the intervening variables is practically meaningless. Under these conditions, for example, one cannot prove on the basis of experimental data whether changes in response strength are the result of an additive or a multiplicative relationship between intervening variables. From more fundamental considerations, Hilgard (1958) points out that Hull's intervening variables cannot in their present form be multiplied meaningfully, since they are not in comparable units of measurement. Certainly the least objectionable formula for reaction potential is also the least specific. Consequently it has the least predictive power:

$$_{S}\overline{E}_{R}=\mathrm{f}(D, K, {_{S}H_{R}}, I_{R}, \text{etc.})$$

In view of the facts here noted, great difficulties arise when Hull and his revisers become more explicit about the nature of the relationships between these variables.

Though it would not be in keeping with the spirit of Hull's formal theorizing, some of the problems might be avoided if Hull's formula for $_{S}\overline{E}_{R}$ were regarded, not as a true mathematical equation, but merely as a kind of shorthand for expressing certain relationships suggested by empirical findings. The arithmetic signs of addition, subtraction, and multiplication in the formula would then not be taken too literally. Thus, $E=H-I$ would not be taken to mean that inhibition subtracts from habit and that when E finally equals zero, the habit has been removed and the organism restored to the same state as before the habit had been acquired. The equation merely states in shorthand form that reaction potential, as inferred from some measure of response strength, decreases as the experimental procedures said to increase habit strength are removed and the conditions said to produce inhibition are applied. The subtraction sign is used here, not in a strict mathematical sense, but only as a shorthand expression for an experimental manipulation. Whether Hull has chosen to add or to multiply various intervening variables most likely has been a result of his attempt primarily to represent known empirical relationships rather than to maintain logical consistency within his theory. He most likely formulated $D\times{_{S}H_{R}}$, for example, because he believed this interaction of habit and drive represented the experimental evidence. And most probably the reason he did not formulate $D\times{_{S}I_{R}}$, even though his theory seems to call for this logically, was simply because he found no evidence that suggests an interaction between drive and inhibition.

From the foregoing considerations, probably the ultimate conclusion to which we are forced regarding the attempted revisions of Hull's theory is not so much that these revisions are no improvement over Hull, but that it is futile to attempt to improve upon Hull by mere juggling of his intervening variables. Hullian theory will not be improved by continuing to work with the concepts of drive, habit, inhibition, etc. in exactly the same form they were given by Hull. The very building blocks of the theory, so to speak, are inadequate, and no amount of recombining them in new ways is likely to result in any substantial advance in learning theory.

Reformulations and Empirical Evidence

${}_sH_R - {}_sI_R$

While Hull (1943) refers to ${}_sI_R$ as a "negative habit," there is no indication in his writing that he regards ${}_sI_R$ as merely negative ${}_sH_R$. The revisions suggested by Osgood and by Jones are based on the assumption that ${}_sH_R$ and ${}_sI_R$ are basically the same phenomenon, ${}_sI_R$ merely being the negative counterpart of ${}_sH_R$. Thus, if they are the same process but merely opposite in effect, it seems logical that one should subtract from the other. Similarly, if ${}_sH_R$ interacts with drive, so should ${}_sI_R$. Hull, however, quite clearly did not regard ${}_sH_R$ and ${}_sI_R$ as basically one and the same phenomenon, and his reasons are based on experimental evidence that reveals differences between the two. Pavlov (1927) originally pointed out the greater susceptibility of internal inhibition (of which ${}_sI_R$ is one variety) to external inhibition (i.e., disinhibition) than is the case with the excitatory process corresponding to Hull's ${}_sH_R$. That ${}_sI_R$ is more labile and sensitive to external influences than is ${}_sH_R$ suggests that it is not merely the negative counterpart of the same phenomenon. Therefore, Hull is consistent with Pavlov in not subtracting ${}_sI_R$ directly from ${}_sH_R$.

Another line of evidence that excitation (conditioning) and inhibition (extinction) are basically different processes is well demonstrated in a series of experiments by Reynolds (1945a, 1945b), which showed that acquisition of a conditioned response is slower for massed than for distributed trials, while the *reverse* relationship holds for extinction. Also a number of studies (Hilgard & Marquis, 1940, p. 119) have shown a *negative* correlation between the speed of conditioning and of extinction.

The issue of whether the generalization gradients of excitation (conditioning) and inhibition (extinction) are the same or different was left undecided by Hull (1943, p. 265). The Bass and Hull (1934) and Hovland (1937) studies referred to by Hull were not adequate to answer this question. Not finding evidence to the contrary, Hull merely assumed that the generalization gradients of excitation and inhibition were the same, which is a convenient assumption in his theory of simple discrimination learning (1943, p. 267) based on the interaction of the gradients of excitation and inhibition. On this point, however, there is now some tentative evidence that seems to contradict Hull's assumption. Liberman (1951) found that extinction $({}_sI_R)$[1] has broader transfer

[1] In Hull's system, though the entire process of extinction is not explained in terms of only ${}_sI_R$, but includes reactive inhibition (I_R) as well, once extinction is complete, or after enough time (probably 5 to 10 minutes) has elapsed for the dissipation of I_R, extinction is conceived of as solely a function of the relative magnitudes of the positive reaction potential and ${}_sI_R$.

effects than acquisition ($_SH_R$). Also there is some evidence (Razran, 1938) that the stimulus generalization of extinction ($_SI_R$) differs from that of excitation ($_SH_R$), in that extinction shows greater stimulus generalization; the gradient of its generalization contains fewer steps; the stimulus generalization of extinction, unlike that of acquisition, does not extend to heterogeneous CRs; and generalization of extinction is more affected by drugs than is generalization of conditioning.

The formulation $_SH_R - _SI_R$ seems misleading in view of the fact that successive periods of acquisition and extinction become more rapid and that an organism in which an acquired response has been extinguished is not the same as an organism that had never acquired the response. Razran (1956) has pointed out that in a partially extinguished CR there can be shown the coexistence of two opposing processes, positive and negative. "Even a wholly extinguished CR bears, by all signs, within itself a two-way CR connection" (p. 42). Successive acquisition and extinction may be conceived of as a kind of discrimination learning, in which both $_SH_R$ and $_SI_R$ grow simultaneously, neither one diminishing the other. The cessation of reinforcement becomes a cue, a conditioned inhibitor, the strength of which increases throughout successive extinction periods (Bullock & Smith, 1953; Perkins & Cacioppo, 1950). This kind of discrimination learning is likely to be a very primitive kind of discrimination not involving symbolic or mediating processes. Tentative evidence for this opinion is found in the experiments on spinal conditioning, which, however, are not yet entirely beyond dispute as examples of true conditioning. Nevertheless, for what it is worth, Shurrager and Shurrager (1946) have reported that both conditioning and extinction, measured at a single synapse in a spinal preparation, become faster with successive periods of conditioning and extinction.

Hull (1952, p. 114) also pointed out that the delay CR (the "inhibition of delay" being due to $_SI_R$) is eliminated by certain drugs, for example, caffeine and benzedrine. It is hard to see why the CR itself would not be markedly weakened or eliminated altogether if these drugs affected both $_SH_R$ and $_SI_R$ in the same manner. The CR is strengthened, however, while the period of delay is markedly shortened. Certain drugs thus seem to have opposite effects on $_SH_R$ and $_SI_R$, suggesting again that they represent essentially different underlying physiological processes. Skinner's (1938, pp. 412–413) finding that benzedrine and caffeine increase the number of responses to a criterion of extinction lends plausibility to the idea that these drugs have different effects on $_SH_R$ and $_SI_R$. If $_SH_R$ and $_SI_R$ were the same process, then a drug increasing $_SH_R$ would also increase the inhibitory effect of each nonreinforced response. If this were the case, the unfailing effect of stimulant drugs in increasing the number of responses to extinction could not easily be accounted for. The evidence bearing on this subject, however, is not crucial, in that we do not have evidence regarding the *percentage* increase in responding during extinction under benzedrine *over* the operant level (preconditioning response rate) under benzedrine. Also it should be noted that the theoretical problem hinges to some extent upon the hypothesized relationship between excitation (or $_SH_R$) and inhibition ($_SI_R$); that is, whether it is the absolute *difference* between the two that matters or the *ratio* (or "balance") between excitation and inhibition. In the Pavlovian system it is the balance or ratio of excitation to inhibition that determines reaction potential. In Hull's system it is the absolute difference between $_{S}H_{R}$ (and the variables interacting with it) and I_R. A strictly Pavlovian revision of Hull might take the following form:

$$_{S}\overline{E}_{R}=\log\frac{D\times{}_{S}H_{R}}{\dot{I}_{R}}$$

Thus it is the balance between excitatory and inhibitory processes that is emphasized and not the absolute difference. In this equation, when the total inhibitory potential ($\dot{I}_R$) is equal in strength to $D\times{}_{S}H_{R}$, the ratio of $D\times{}_{S}H_{R}/I_{R}$ becomes 1.0, and since log 1.0 = 0, the effective reaction potential ($_{S}\overline{E}_{R}$) will equal zero.

The fact that Eysenck and his coworkers have subscribed to the Jones revision would seem incompatible with Eysenck's (1956) theory concerning the *extinction* of $_{S}I_{R}$. The *extinction* of $_{S}I_{R}$ is paradoxical and inconsistent with other aspects of Hull's theory and also of Jones' revision. If, as maintained by Jones and by Eysenck, $_{S}I_{R}$ is merely negative $_{S}H_{R}$, then the mere lack of reinforcement of $_{S}I_{R}$ (reinforcement being the dissipation or avoidance of I_R) should not result in a decrease in $_{S}I_{R}$. Lack of reinforcement does not diminish the $_{S}H_{R}$ already present, so it should not diminish $_{S}I_{R}$ either. The notion that extinction is an active process of an increasing inhibition ($\dot{I}_R$) depressing performance ($_{S}E_{R}$) is basic in Hull's system. It, therefore, seems absurd, while remaining in the Hullian framework, to speak of the extinction of inhibition without first postulating a second inhibitory process which depresses the first. Fortunately, there is no experimental evidence at present to suggest that such a complication would be necessary.

$D\times{}_{S}I_{R}$

In Hull's theory there is no interaction between drive and conditioned inhibition. The $D\times{}_{S}I_{R}$ interaction, however, is explicit in a number of the revisions. Since $_{S}I_{R}$ is the primary and essential intervening variable accounting for experimental extinction, we may well examine the different predictions generated by Hull and the revisions with respect to the $D\times{}_{S}I_{R}$ interaction.

According to Hull, since D multiplies only $_{S}H_{R}$ and not $_{S}I_{R}$, we should predict that certain measures of extinction will be affected by changes in D. With the Hullian formula $D\times{}_{S}H_{R}-{}_{S}I_{R}$, one can predict that under a high drive level there will be a greater number of responses to extinction (n) than under low drive. The same increment of $_{S}I_{R}$ is generated by each response during extinction, regardless of the level of D, while the positive reaction potential ($D\times{}_{S}H_{R}$) is increased by a higher level of D. Not only does it follow from Hull's formula that a greater number of responses is required for extinction, but extinction curves under high and low D should be parallel. They approach the criterion of extinction with the *same slope*, but reach it at *different points*.

The revisions containing the $D\times{}_{S}I_{R}$ interaction generate predictions that are exactly opposite to the foregoing. If net reaction potential is a resultant of $D\times{}_{S}H_{R}-D\times{}_{S}I_{R}$, then every increment of $_{S}I_{R}$ will be increased by D to the same degree that $_{S}H_{R}$ has been increased. Consequently, there should be the

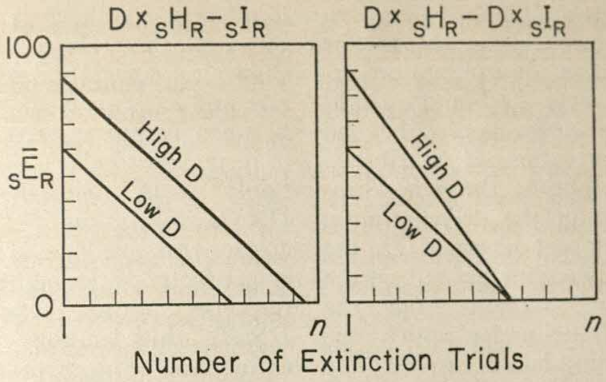


FIG. 1. The relationships between drive (D), number of trials to extinction (n), and effective reaction potential ($_S\bar{E}_R$) as predicted by Hull's formulation (left) and by Jones's formulation (right).

same number of responses to extinction under high drive as under low drive. Also, the slopes of the extinction curves, as measured by, say, rate of responding, would be different under high and low drive. In other words, the curves would approach the criterion of extinction with *different slopes*, but would reach it at the *same point*.

If the proponents of the $D \times {}_SI_R$ formulation object to the foregoing predictions on the grounds that I_R has not been taken into account, let it be pointed out that $_SI_R$ is essential for complete extinction of the response and that extinction can take place with sufficiently spaced trials to prevent the growth of I_R. If, as Hull hypothesized (1943, pp. 300–301), the formation of $_SI_R$ is dependent upon nonresponding being coincident with the dissipation of I_R, extinction could not take place if all I_R had dissipated in the interval between each presentation of the nonreinforced CS. Yet extinction is known to occur even with long intertrial intervals of 24 hours or more, when I_R should supposedly have been completely dissipated (Razran, 1956, p. 43). This, along with the fact that in all of the revisions an increment of I_R will reduce $_SE_R$ by the same proportion regardless of the level of D, makes I_R irrelevant to the present argument. (The $D-I_R$ formulation is discussed at a later point.)

There is a considerable amount of experimental evidence bearing on the above predictions. The preponderance of evidence favors the Hullian formula and fails to support the notion of a $D \times {}_SI_R$ interaction. Perin (1942), working with rats, found a marked positive relationship between D (degree of hunger) at the time of extinction and the number of responses required for extinction. Brandauer (1953) extinguished bar pressing in rats under three levels of drive (thirst) and found a positive relationship between strength of drive and number of responses during extinction. Even under minimal differences in hunger drive (.5, 1, 2 hours' deprivation) Saltzman and Koch (1948) found highly significant differences in number of responses to extinction in a modified Skinner box. Brown (1956) also found that rats on high drive make more responses during extinction than those on low drive. Cautela (1956) showed essentially the same relationship for the extinction of a discrimination response. However, he found a slight decrease in *n* for levels of *D* beyond 23 hours' deprivation. He attributed this phenomenon to the generalization gradient of the drive stimuli; under the highest levels of *D*, the drive stimuli were further out on the generalization gradient from the drive conditions under which the original learning had occurred. The energizing and stimulus properties of drive are thus apt to interact in this type of experiment.

In experiments with human subjects, where anxiety has been used as a measure of drive, a similar relationship with extinction has been found. In one study, high anxiety subjects required almost twice the number of trials to extinguish the conditioned eyeblink as did low anxiety subjects (Spence & Farber, 1953). Bitterman and Holtzman (1952) obtained similar results in extinguishing the PGR in high and low anxiety subjects.

Skinner's (1938) early notion of the "reflex reserve" appears to be consistent with the $D \times {}_SI_R$ formulation. Skinner believed that the number of responses emitted during extinction was solely a function of the number of previously reinforced responses and the schedule of reinforcement. Thus drive should not affect *n*, but would affect only the *rate* of emission of responses. The reflex reserve concept, however, has long since been found unfruitful. While theoretically it is probably not a strictly testable hypothesis, it now at least appears quite incorrect in view of the evidence (Ellson, 1939). Skinner's (1938) original belief that *rate* of responding, but not the number of responses in extinction (*n*), is affected by drive is contradicted by Bullock's (1950) investigation showing a correlation of .61 between rate and *n*. This positive correlation between response rate and number of responses to extinction would certainly seem inconsistent with a $D \times {}_SI_R$ formulation. If drive increases response rate, ${}_SI_R$ should increase faster under higher drive, each response adding the increment $D \times {}_SI_R$, thus leading to more rapid extinction. The evidence is exactly the contrary. Higher drive not only increases the rate of response, but also increases the total number of responses to a criterion of extinction.

The best available evidence also indicates that the slope of the extinction curve is the same under high and low drive, as would be predicted from Hull's theory. Sackett (1939) showed that when the extinction curves of two groups of rats, one group extinguished under 6 hours' hunger drive and the other under 30 hours' drive, are Vincentized, the forms of the two curves are almost identical. The 30-hour group produced more responses to extinction and required more time to extinguish, but the slope of the extinction curve was the same as that of the 6-hour group. Barry (1958) trained rats in a running response and extinguished them under high and low drive. The extinction curves were parallel, and when drive was equalized in both groups late in extinction, the curves converged and were identical after three trials. When drive was equal for both groups early in extinction, and then, later in extinction, the groups were run under high and low drive, the extinction curves diverged, and, after three trials, continued almost parallel, as would be predicted from Hull. (The fact that it took three trials, rather than one, for the curves to converge or diverge after the change in D, however, is somewhat embarrassing to Hull's theory as it is also to the revision.) Both these findings are consistent with the $D \times {}_SH_R - {}_SI_R$ formulation and not with $D \times {}_SH_R - D \times {}_SI_R$. But these experiments cannot be regarded as at all definitive in view of the finding of Reynolds, Marx, and Henderson (1952) of an interaction between D and the incentive factor K (a function of amount of reward). This interaction plays havoc with any theoretical conclusions drawn from experiments on the effects of drive on extinction in which the incentive factor has not been taken into account. Reynolds et al. (1952) had four groups of rats learn bar pressing under all combinations of high drive–low drive and large reward–small reward. All animals were given extinction trials under equal drive. It was found that

> in those learning situations where a relatively large amount of reward is employed for reinforcement, high D animals extinguish more readily than low D animals; and . . . where a relatively small reward is given per reinforcement, low D animals extinguish more readily than high D animals (pp. 41–42).

Hull's theory and its revisions generate conflicting predictions regarding spontaneous recovery. In the Jones (1958) formula, ${}_S\bar{E}_R = D - I_R) \times ({}_SH_R - {}_SI_R)$, spontaneous recovery could occur only if at the end of the first set of extinction trials $D - I_R = 0$. But this formulation would lead to problems, since, if $D - I_R = 0$, no habits at all could be activated temporarily until some of the I_R had dissipated, and no behavior of any kind would occur after the end of the first extinction period. We know very well, however, that animals go on behaving in various ways immediately following the extinction of a particular response. But then if we do not wish to assume that $D - I_R$ is equal to zero immediately after the first extinction period, we must assume that ${}_SH_R - {}_SI_R$ equals zero, or extinction would not have occurred. Yet if ${}_SH_R - {}_SI_R$ were zero, there could be no spontaneous recovery. Conceivably one way out of this dilemma for the Jones revision is to make some assumptions about a reaction threshold which must be exceeded before an overt response is made. Thus, overt extinction could occur before either $D - I_R = 0$ or ${}_SH_R - {}_SI_R = 0$. Spontaneous recovery would then result from the dissipation of I_R, as in Hull's theory. If this were true, one might predict from the Jones revision that there would be very little, if any, spontaneous recovery after extinction under high drive, but greater amounts of spontaneous recovery after extinction under low drive. Since $D - I_R$ would approach the threshold value quickly where D is initially low, there would result an appreciable increase in D, and hence of response strength, with the dissipation of I_R, and spontaneous recovery would result. Under high drive $D - I_R$ would not approach the threshold value as quickly as would ${}_SH_R - {}_SI_R$. Thus, since ${}_SH_R - {}_SI_R$ would be a smaller value after the first extinction, there should be less spontaneous recovery at the beginning of subsequent extinction periods.

Different predictions may be made from Hull and the $D \times {}_SI_R$ revision concerning the effect of an increase in drive after extinction is complete. According to Hull's $(D \times {}_SH_R) - {}_SI_R$, an increase in drive after complete extinction should result in further "spontaneous recovery." According to the $D \times ({}_SH_R - {}_SI_R)$ formulation, once extinction is complete (i.e.,

${}_{S}H_{R}-{}_{S}I_{R}=0$), an increase in D should not produce any "spontaneous recovery."

Unfortunately, the experimental evidence bearing on all these predictions is meagre, conflicting, and inconclusive. Hull (1943, p. 249) cites Pavlov's finding that an increase in drive after extinction is complete causes the reappearance of the CR in the presence of the CS. This is, of course, consistent with Hull's formulation, but not with the $D \times {}_{S}I_{R}$ formulation. The same phenomenon seems to occur also in instrumental conditioning. Jenkins and Daugherty (1951) extinguished a pecking response in pigeons under three levels of drive. They found that the number of responses in extinction is a function of drive level and that when extinction was relatively complete an increase in drive caused gross recovery of the conditioned behavior. The authors used the term "relatively complete" extinction because the pecking response in pigeons never seems to be completely extinguished. But the recovery of a "relatively extinguished" CR under increased drive is certainly more consistent with $(D \times {}_{S}H_{R})-{}_{S}I_{R}$ than with $D \times ({}_{S}H_{R}-{}_{S}I_{R})$. The writer knows of only one study that appears to contradict the finding of Jenkins and Daugherty. Crocetti (1952) found that when rats were "completely" extinguished in a Skinner box, increase in drive did not increase the response rate over the preconditioning response rate under the higher level of drive. (Extinction was considered complete when the response rate became equal to the operant level prior to conditioning.) This finding is, of course, inconsistent with Hull's $(D \times {}_{S}H_{R})-{}_{S}I_{R}$. Crocetti did not control for the changes in the drive stimulus (S_D) with increased hunger, and so his finding is not definitive with respect to the present theoretical issue. If we assume that ${}_{S}H_{R}$ and ${}_{S}I_{R}$ are conditioned to S_D as well as to other stimuli, then the changes in S_D from the conditioning trials to the extinction trials or spontaneous recovery trials becomes a crucial point in this type of experiment. Fortunately in an experiment by Lewis and Cotton (1957) the effect of such changes in S_D was taken into account. Three groups of rats were trained in a running response under three levels of drive, viz., 1, 6, and 22 hours' food deprivation. Each group was then divided into three groups which underwent extinction under 1, 6, and 22 hours' drive. Extinction proceeded more rapidly under lower drive, as would be expected from Hull's formulation, but drive level seemed to have no effect on the magnitude of spontaneous recovery, a fact which is inconsistent with $(D \times {}_{S}H_{R})-{}_{S}I_{R}$. But the $D \times ({}_{S}H_{R}-{}_{S}I_{R})$ revision cannot comprehend both of these findings either, for with this formulation drive level should have no effect on number of trials to a criterion of extinction. It seems obvious that clarification of the effects of drive on spontaneous recovery must await further experimentation which is specifically designed for this purpose and which takes into account both the energizing and stimulus properties of drive. Some of the lack of consistency and agreement in this area may also be due to interspecies differences and to the use of different measures of response strength. Latency, running time, response rate, and number of trials to extinction are used singly in different studies as measures of response strength even though they are far from being perfectly correlated. Each measure undoubtedly involves certain parameters peculiar to itself. To use only one such measure of response strength and only one species of animal is an inadequate method for testing a precise deduction from a general behavior theory.

In the delayed CR, the inhibition of the response during the period of delay is attributed in the Hullian system to ${}_{S}I_{R}$ (Hull, 1952, p. 114). Consistent with Hull's formulation of $D \times {}_{S}H_{R} - {}_{S}I_{R}$ is the fact that an increase in D lessens or eliminates the period of delay in the CR. The $D \times {}_{S}I_{R}$ formulation does not accommodate this fact, but leads to an opposite prediction, i.e., an increase in D should strengthen the inhibition of delay.

One of the weakest points in Hull's system involves the dependence of ${}_{S}I_{R}$ upon I_{R}. It is no less troublesome to any of the revisions. (Spence excepted, since his inhibition concept has nothing in common with I_{R}.) It is stated that I_{R} is generated whenever a response is made, the amount of I_{R} being a function of the effortfulness of the response, and that I_{R} rapidly dissipates, accumulating only if responses follow one another in rapid succession. The dissipation of I_{R}, a "negative drive state," reinforces the habit of not responding, which is ${}_{S}I_{R}$. This hypothesis encounters obvious difficulties. If a response is followed by the dissipation of I_{R}, this would seem to have all the requirements for reinforcing the response, leading to increased response strength rather than extinction.[2] Also, subzero extinction would be unlikely if increases in ${}_{S}I_{R}$ were dependent upon *reactive* inhibition (I_{R}). And it is almost impossible to explain the extinction of relatively effortless CRs, such as salivation, eyeblink, and the alpha rhythm, when the extinction trials are widely spaced. Pavlov (1927, p. 76) obtained rapid extinction of the salivary CR using only one presentation of the CS per day. Razran (1956, p. 43) has reviewed the evidence that contradicts a theory of extinction based on reactive inhibition. There are even cases where spaced trials have led to more rapid exintction than massed trials (Sheffield, 1950; Stanley, 1952). Kimble (1950) has argued from studies of motor learning that a certain threshold or critical level of I_{R} must be reached before ${}_{S}I_{R}$ develops. Motor learning experiments have presumably shown that I_{R} can form without leaving behind any ${}_{S}I_{R}$. This is inconsistent with extinction based on widely spaced trials. In fact, it does not seem to the writer that the Hullian inhibition postulates, as they have been used in the field of motor learning, represent the same processes found in extinction phenomena. It has been a case of giving the same theoretical labels to basically different processes. The most fundamental difference between ${}_{S}I_{R}$ in conditioning and in motor learning has to do with the amount of response necessary to produce ${}_{S}I_{R}$. Five or six minutes of pursuit rotor practice seems necessary before ${}_{S}I_{R}$ is in evi-

[2] One can get around this problem, of course, by invoking the gradient of reinforcement. If the time required for the dissipation of I_{R} is greater than the effective gradient of reinforcement, the foregoing proposition would not hold true. At present there is no basis for arguing the point. While Hull gives 20–30 seconds as the maximum delay between the response and reinforcement if reinforcement is to be effective, the time required for the dissipation of I_{R} is solely a function of the amount of I_{R} generated by the response and, therefore, is variable, although the *rate* of dissipation of I_{R} may not be variable. Perhaps an even simpler way out is the idea that I_{R} leads to a "resting response" which in turn is reinforced by the dissipation of I_{R}.

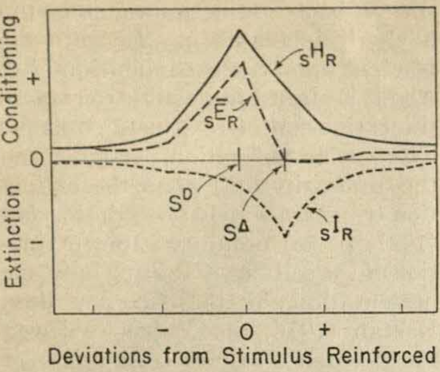


FIG. 2. Illustrates algebraic summation theory of discrimination. (Effective reaction potential, $_s\bar{E}_R$, is a result of subtraction of generalized extinction, $_sI_R$, from generalized conditioning, $_sH_R$. See text for full explanation.)

dence, while only a single conditioned response, such as salivation, PGR, or eyeblink, is evidently sufficient to produce $_sI_R$. Thus it does not seem that the $_sI_R$ invoked in theories of motor learning could be the same $_sI_R$ as that in Hull's theory of conditioning.

It is also held by Hull, and even more explicitly by his revisers, that the amount of $_sI_R$ built up per trial is related to the amount of I_R dissipated, the dissipation acting as reinforcement for the negative habit, $_sI_R$. But this is inconsistent with Hull's own revision of his theory (Hull, 1951), in which the growth of habit, $_sH_R$, and presumably also of negative habit, $_sI_R$, is a function only of the *number* of reinforcements and not the *amount* of reinforcement. None of these awkward predicaments has been remedied by the revisions here reviewed. Those revisions insisting on the theoretical equivalence of $_sH_R$ and $_sI_R$ as being merely positive and negative habits have retained one of the weakest elements in the Hullian system.

Discrimination learning. If discrimination learning involves an increase in habit strength to the positive stimulus (S^D) and an increase in inhibition (I_R and $_sI_R$) to the negative stimulus (S^Δ), then the effects of drive on discrimination learning should be highly germane to the plausibility of the $D \times {_sI_R}$ formulation. Jones (1958) invokes Spence's (1937) theory of discrimination learning, adapted by Hull, involving the overlapping generalization gradients of $_sH_R$ and $_sI_R$, in support of the $D \times {_sI_R}$ part of his revision. This theory is illustrated in Figure 2. The discrimination would be perfect (except for behavioral oscillation) when the net reaction potential resulting from subtracting the generalized $_sI_R$ from the generalized $_sH_R$ is some positive value for S^D and zero for S^Δ, as in Figure 2. Jones (1958) argues that, according to Hull's $D \times {_sH_R} - {_sI_R}$, an increase in D should obliterate the learned discrimination. Since some discriminations are not obliterated or even weakened by an increase in D, Jones reasons that $_sI_R$ must also be multiplied by D, so that the increase in $_sI_R$ will be proportional to the increase of $_sH_R$ when multiplied by D, thereby preserving the discrimination.

Before Jones' argument can be evaluated, some clarification of the Spence-Hull theory of discrimination learning is necessary. In the first place, there is often confusion concerning whether discrimination learning is a matter only of the *relative* strengths of $_s\bar{E}_R$ to the S^D and S^Δ, or whether the formation of a discrimination requires the reduction of $_s\bar{E}_R$ to the S^Δ to zero or at least below the operant level of the response, i.e., below the strength of the response before any conditioning or extinction has occurred. If the former, then all that would be necessary for discrimination to take place would be that the S^D have greater $_S\overline{E}_R$ than the S^Δ. The $_S\overline{E}_R$ to the S^Δ would not necessarily have to undergo some degree of extinction. If this were the case, Jones' use of the Spence-Hull theory of discrimination, as illustrated in Figure 2, would not be applicable to the present argument concerning the effects of drive on discrimination learning. The evidence, however, strongly suggests that the $_S\overline{E}_R$ to the S^Δ must undergo some degree of extinction for discrimination to become nearly perfect. To this extent, at least, the Spence-Hull theory appears to be correct.

For example, Grice (1948) gave one group of rats 200 rewarded trials in responding to a disc 8 centimeters in diameter and gave another group of rats 200 rewarded trials in responding to a 5-centimeter disc. Then both groups were given discrimination training, with the 8-centimeter disc as the S^D and the 5-centimeter disc as the S^Δ. The group which had been previously rewarded on the 8-centimeter disc learned the discrimination faster. Now if all that were involved in discrimination were the *relative* response strengths to the S^D and S^Δ, the 8-centimeter group should have learned to make the discrimination immediately, since response to the S^D had already been rewarded on 200 trials, and the response strength to the S^Δ resulting from stimulus generalization would have been less than the response strength to the S^D. Since the learning curve for the acquisition of the discrimination is very gradual, however, it suggests that *extinction* of the response to the S^Δ through nonreinforcement is necessary for the learning of the discrimination.

An even more cogent demonstration of the necessity for extinction of S^Δ in discrimination learning is an experiment by Fitzwater (1952). Three groups of rats were used: Groups A, B, and C. In preliminary training Group A was run an equal number of times into each of two alleys having differential cues—call them *X* and *Y*, respectively. *X* was always reinforced; *Y* was never reinforced. Group B was run an equal number of times into each of two alleys having the Cues *X* and *Z*. *X* was always reinforced; *Z* was never reinforced. Group C was run only into one alley, with Cue *X*, the same number of times as the other groups. Then discrimination training began, with the animals having to learn to discriminate *X* as the S^D and *Y* as the S^Δ. Group A learned the discrimination most rapidly, while Groups B and C did not differ significantly in speed of learning. The theoretical interpretation of these results is that Group A had already built up inhibition to the S^Δ, while Groups B and C had not. Fitzwater concluded that

> apparently in a visual discrimination task it is about as important to establish an avoidance habit as an approach habit, and that an appreciable discrimination does not seem to occur if an approach habit is established alone (p. 480).

The terms "approach habit" and "avoidance habit" may be interpreted in the context of the present discussion as excitation (or $_SH_R$) and extinction ($_SI_R$), respectively. Thus it is apparent that a decrease in $_S\overline{E}_R$ to the S^Δ as well as an increase in $_S\overline{E}_R$ to the S^D is necessary for discrimination learning. It is not just a matter of $_S\overline{E}_R$ to the S^D being relatively greater than to the S^Δ.

Another experiment by Grice (1949) offers further evidence that discrimination depends upon the *extinction* of the response to the S^Δ and not merely a relative difference in response strengths between S^D and S^Δ. One group of rats was trained in a visual size discrimination with S^D and S^Δ presented simultaneously, and another group was trained on the same discrimination with S^D and S^Δ presented successively in random order. Grice found no difference between the "simultaneous" and "successive" groups in the rate of learning the discrimination. In both cases learning apparently consisted of increasing the response strength to the S^D and completely extinguishing the response to S^Δ. Furthermore it was found that the rats which learned the problem as a pair (i.e., simultaneous presentation) responded differently to the S^D and S^Δ when they appeared singly, showing that even under simultaneous presentation of the S^D and S^Δ, the response to the S^Δ had undergone extinction.

It is not maintained that *complete* extinction of the response to S^Δ is necessary. Extinction is a relative matter and is probably best measured, not in relation to some theoretical "absolute zero," but in relation to the "operant level" or probability of occurrence of the particular response before extinction trials have been assumed to take place. In the Grice (1949) experiment there was a decrease in latency of response to S^D and an increase in latency of response to S^Δ, whether the two stimuli were presented simultaneously or successively. $_{S}\bar{E}_{R}$ to the S^Δ, as measured by latency, was considerably less at the end of discrimination training than at the beginning. In fact, extinction of response to S^Δ may play a greater role in discrimination learning than does the strengthening of the response to S^D. Webb (1950) trained rats to jump to a black-white discrimination until it was well learned. When, after training, only the S^D was presented to the rats, the mean latency of their response was 2.0 seconds, which was not significantly less than the pretraining latency. On the other hand, when only the S^Δ was presented, the mean latency of response was 80.5 seconds, which may be interpreted as indicating considerable extinction or inhibition of the response to the S^Δ. If one defines the zero level of $_{S}\bar{E}_{R}$ in the Hull-Spence model in Figure 2 simply as the operant level (i.e., the pretraining latency or probability of responding to the particular stimuli), then this model appears to be quite consistent with the experimental evidence in showing that discrimination depends upon extinction of the response to the S^Δ.

This model, however, seems to be deficient in some other respects. Hanson (1957), for example, performed a very careful experiment which led to the conclusion that over-all response strength is *not* weakened by discrimination training, as would be predicted from the Spence-Hull model. (That is, since the resultant $_{S}\bar{E}_{R}$ is the algebraic sum of generalized excitation and inhibition, $_{S}\bar{E}_{R}$ to the S^D should be less after discrimination training than it would be in simple conditioning to a single stimulus.) Hanson concluded that

> the major result of discrimination training is to bring a large proportion of the responses available in extinction under the control of another range of stimuli, those which do not ordinarily gain control of the response as the result of simple conditioning without differential reinforcement (p. 889).

This conclusion is not compatible with the Spence-Hull theory.

It may be argued that Jones has taken the Spence-Hull diagram (Figure 2) too literally. Very little is known about the actual shapes of the generalization gradients of $_SH_R$ and $_SI_R$, and until a proper metric is worked out, arguments over this point cannot be settled. What little evidence there is, though far from conclusive, suggests that the generalization gradients of excitation and inhibition are probably different in a number of respects (Razran, 1938). Furthermore, the amount of overlap of the gradients of excitation and inhibition will depend on the distance apart of S^D and S^Δ, and there is reason to believe that the effects of drive on discrimination will interact with the degree of disparity between S^D and S^Δ (Broadhurst, 1957). We would predict from Hull's $D \times {_SH_R} - {_SI_R}$ that the farther apart S^D and S^Δ are, the less deleterious to the discrimination are the effects of increased drive. This essentially is the Yerkes-Dodson Law (Yerkes & Dodson, 1908), which, in its most general form, states that the optimum motivation for a learning task decreases with increasing difficulty. This relationship between drive and difficulty of discrimination, however, cannot be predicted from the Jones formulation of $D \times ({_SH_R} - {_SI_R})$.

Rather than arguing from a highly hypothetical model involving the relative shapes and magnitudes of the generalization gradients of $_SH_R$ and $_SI_R$, as Jones has done, we can better make predictions concerning the directly observable effects of increased drive on discriminations. What is the effect of drive on the initial learning of a discrimination, and does an increase in drive have a different effect on the learning of easy and difficult discriminations, as determined by time required for learning? What is the effect of change in drive on discriminations that are already established? What effect does a change in drive have on the extinction of a discrimination?

In discrimination learning, since the relative amounts of $_SH_R$ and $_SI_R$ built up to the S^D and S^Δ are different, we would expect from the $D \times ({_SH_R} - {_SI_R})$ formula that an increase in D would always have a facilitative effect on learning a discrimination. The degree of facilitation would depend upon the degree of difference between S^D and S^Δ. If we assumed considerable overlapping of generalization gradients, then there would be relatively little effect of an increase in D. If the discrimination were easy, increases in D should improve the discrimination, since the relatively greater $_SH_R$ to the S_D and the relatively greater $_SI_R$ to the S^Δ would both he multiplied by D. In no case should discrimination be weakened by an increase in D.

On the other hand, if we assume that response to S^Δ must undergo extinction for a discrimination to be learned, Hull's formula $D \times {_SH_R} - {_SI_R}$ leads to quite different predictions, viz., that increase in D should weaken difficult discriminations, where one might assume overlap of the stimulus generalization gradients, but would strengthen discriminations in which S^D and S^Δ are widely separated on the generalization gradient.

What is the evidence? We have already mentioned the Yerkes-Dodson Law, which is possibly consistent with Hull, but certainly not with the $D \times ({_SH_R} - {_SI_R})$ formula. Broadhurst (1957) has demonstrated this "law" most effectively, using rats in a brightness discrimination problem and manipulating drive by means of oxygen deprivation. Skinner (1938, p. 188) has observed that it is important in establishing discriminant operant conditioning to keep the hunger drive as constant as possible, for changes in drive disturb the discrimination. More explicitly, Teel (1952) has shown that in selective learning, in which correct responses are reinforced and incorrect responses are nonreinforced or extinguished, rats under high drive (food deprivation) require a *greater* number of trials to reach a criterion of learning than rats under low drive. One cannot predict these facts from the $D\times({}_sH_R-{}_sI_R)$ formula. In fact, just the opposite outcome would be predicted for the Teel experiment. With human subjects, Hilgard, Jones, and Kaplan (1951) found that high anxiety subjects (on Taylor Manifest Anxiety scale) had greater difficulty than low anxiety subjects in forming a *discriminatory* CR. It is well-established that anxious subjects develop simple eyeblink CRs more readily than nonanxious subjects. (This relationship has not been found to hold, however, for autonomic CRs.) Interpreting anxiety as a drive, both sets of findings are consistent with Hull, but not with $D\times({}_sH_R-{}_sI_R)$. An experiment by Spence and Farber (1954) found that the difference between high and low anxious subjects in forming a discriminatory response showed up only on the S^D but not on the S^Δ. That is, D (anxiety) seemed to affect only the CS (i.e., S^D) associated with relatively greater ${}_sH_R$ and not the CS (i.e., S^Δ) associated with relatively greater ${}_sI_R$. Spence interprets this finding as evidence that D interacts with excitation (${}_sH_R$) but not with inhibition (${}_sI_R$).

In a well-established discrimination, in which S^D and S^Δ are relatively far apart on the stimulus generalization gradient, and in which relatively more ${}_sH_R$ than ${}_sI_R$ has been built up to S^D than to S^Δ, and relatively more ${}_sI_R$ built up to S^Δ than to S^D, we would predict from $D\times({}_sH_R-{}_sI_R)$ an improvement in the discrimination with an increase in drive. That is, the ratio of number of responses to S^D to number of responses to S^Δ should increase, since response to S^D is increased by $D\times{}_sH_R$, and inhibition of response to S^Δ is increased by $D\times{}_sI_R$. Dinsmoor (1952) performed an experiment bearing on this point. A simple discrimination habit was well-established in rats in the Skinner box, with S^D being the presence of light and S^Δ being total darkness. When D was increased to varying degrees by food deprivation, the *number* of responses per unit of time to both S^D and S^Δ increased, but the *ratio* of S^D and S^Δ responses remained exactly the same at seven different degrees of hunger. In short, the discrimination was not improved by an increase in D. Though Hull's theory is not sufficiently quantified to have precisely predicted the outcome of this particular experiment (because absolute levels of D and ${}_sH_R$ as well as the jnd's between S^D and S^Δ must be taken into account), at least the result is consistent with the $(D\times{}_sH_R)-{}_sI_R$ formulation.

There is other experimental evidence, however, which suggests that *both* the Hullian and the revised formulations are inadequate to explain the effects of drive on discrimination learning. A number of studies have found no relationship at all between drive and proficiency in selective learning or solving discrimination problems (Meyer, 1951; Miles, 1959; and a number of doctoral dissertations reported by Spence, Goodrich, & Ross, 1959). Spence et al. (1959) have scrutinized the conflicting findings in this field with a view to discovering the reason for the lack of agreement between various investigations on the effect of drive on selective learning and discrimination. They arrived at the hypothesis that performance in selective learning (such as learning a black-white discrimination) is independent of drive level when responses to the S^D and S^Δ are equated, but varies with drive when responses are not equated. They performed a set of experiments which supported this hypothesis. The results are inexplicable in terms of Hull's theory or any of its revisions except that of Spence. These findings suggest that the growth of $_SH_R$ is not a function of number of reinforced responses, as in Hull's system, but is a function merely of the number of responses, whether reinforced or not. The growth of inhibition is a function only of the number of nonreinforced trials. This formulation will account for the major finding of the experiment by Spence et al. (1959). But another aspect of their findings remains inexplicable in terms of any current theory of learning. When responses to S^D and S^Δ were equated, an increase in drive increased the response strength to *both* the S^D and S^Δ. But when the rats were forced to respond twice as often to S^D as to S^Δ, an increase in drive *increased* the response strength to S^D but *decreased* response strength to S^Δ. Spence et al. concluded that

> the results of the two (experiments) are in fundamental disagreement so far as the effects of drive differences on the strength of nonreinforced responses are concerned. It is perhaps obvious that we need to obtain much more knowledge than we now possess concerning the variables affecting the development of response decrement with nonreinforcement. Unfortunately, this problem has been badly neglected in conditioning experiments with the consequence that such an empirically based theory as the present one [i.e., Spence's theory] is weakest in this area (p. 15).

Though the present state of our knowledge in this area does not permit any definite conclusion regarding the effects of drive on discrimination learning, it appears that no current theory is able to comprehend all the relevant facts now available.

But now let us ask: What happens when a discrimination is *extinguished* under various levels of drive? Cautela (1956) trained rats in a discrimination under 23 hours' food deprivation and then extinguished the discriminative response under 0, 6, 12, 23, 47, and 71 hours' deprivation. The criterion of extinction was failure to respond to either S^D or S^Δ within 3 minutes. Many more responses were required for extinction under high drive levels (23, 47, or 71 hours' deprivation) than under low drive (0, 6, or 12 hours). This result can be predicted from $D \times {_SH_R} - {_SI_R}$. On the other hand, it is difficult to see why a change in drive should have any effect on the number of responses to extinction if $_SH_R$ and $_SI_R$ are *both* increased or decreased proportionately by changes in D, as stated in the revised formula.

$D - I_R$

Since Hull referred to reactive inhibition (I_R) as a "negative drive," he has been accused of logical inconsistency for adding a drive to a habit (i.e., $I_R + {_SI_R}$) and the suggested remedy has been the obvious one, viz., to subtract I_R from D. But predictions from this formulation lead to empirical embarrassment. For example, when extinction is carried out under massed trials, and, after a period of rest, there is some spontaneous recover, we must assume, according to the $_S\bar{E}_R = (D - I_R) \times ({_SH_R} - {_SI_R})$ formulation, that $D - I_R = 0$ at the end of the first extinction period. For there would be no spontaneous recovery if it were $_SH_R - {_SI_R}$

that had become equal to zero. Yet, according to Hullian theory (including the revisions), no behavior can occur unless D is greater than zero. And it is known that an animal at the end of extinction is far from being inactive. Only the extinguished CR becomes inactive, while other behavior in the animal's repertoire is immediately evident. Theoretically this could not be so if the drive component in the equation for reaction potential were zero.

Experimental evidence contradicting $D-I_R$ is presented by Hull (1952, p. 50). A rat is trained to press either of two bars in different locations in a Skinner box to obtain food. During extinction the rat alternates its response from one bar to the other. I_R does not have to dissipate before the alternate bar can be pressed. This strongly suggests that I_R must be associated with the particular response, rather than cause a diminution in the total drive state, which in the Hullian system is an amalgam of all the organic needs of the moment and their associated "drive stimuli" (S_D).

In an experiment highly relevant to this point, Smith and Hay (1954) took advantage of the great sensitivity to changes in drive of rate of responding in the Skinner box. As soon as operant conditioning had led to a stable response rate, a discriminatory stimulus was introduced, the S^D always being reinforced, the S^Δ never. During the learning of the discrimination, the number of responses to S^D increased while the number of responses to S^Δ decreased, but the *rate of responding remained constant.* If the extinction of S^Δ had involved $D-I_R$, there should have been the decrease in over-all response rate which is associated with lowered drive. On the other hand, this finding is entirely consistent with Hull's formulation.

$I_R \times {}_SI_R$

Here we have a formulation which, if the rules of algebra are followed religiously in manipulating Hullian variables, leads to a paradox—a positive addition to reaction potential resulting from the interaction of two inhibitory variables. Jones (1958) even goes on to say that the paradoxical outcome of $I_R \times {}_SI_R$ increasing ${}_S\overline{E}_R$ might explain the "ultraparadoxical effect" described by Pavlov (1927). This might be called explanation by clang association.[3] It is difficult for the writer to understand why Jones and other revisers have so gratuitously regarded the minus sign as being permanently attached to I_R and ${}_SI_R$. Though these quantities are subtracted from positive reaction potential, the negative sign is not necessarily an inherent part of these inhibition variables. Even if I_R and ${}_SI_R$ were multiplicatively related, there is no reason why their product could not be subtracted from the positive reaction potential.

The empirical evidence regarding the $I_R \times {}_SI_R$ interaction is far from satisfactory, for there is always an "out" via the possible interaction of all the other intervening variables in the system. But in terms of sheer plausibility—and that is all one can

[3] The "paradoxical" and "ultraparadoxical" effects observed by Pavlov, in which a weaker intensity of the CS will elicit a CR that had been extinguished to a stronger intensity of the CS, are probably best explained in terms of a generalization gradient on the stimulus intensity dimension. Because of the gradient, extinctive inhibition built up to a CS of one intensity will not be sufficient to inhibit the CR to a CS of a different intensity, even though it be weaker. Or the effect may be explained as disinhibition caused by a "novel" stimulus—novel because the intensity is weaker than that of the original CS.

go on at present—it must be said that $I_R \times {}_sI_R$ is a weak formulation. The only relevant evidence comes from experiments on motor learning, the one area in which there are rather clear-cut operational definitions of what constitutes I_R and ${}_sI_R$. In general, performance decrement that dissipates during rest is identified with I_R; the decrement that still remains after rest is identified with ${}_sI_R$.

Duncan (1951) gave two groups of subjects massed and distributed practice on the pursuit rotor. During this 5-minute practice period, the massed group presumably would develop more I_R and hence more ${}_sI_R$. Then both groups were allowed 10 minutes of rest, so that at the beginning of the postrest trials, nearly all I_R should have dissipated, leaving the two groups differing only in ${}_sI_R$. The postrest trials were massed for both groups. Here exist the very conditions which should allow an $I_R \times {}_sI_R$ interaction to show itself. If there were an interaction, the postrest performance curves of the two groups should diverge. In fact, they did not diverge, or converge, but ran exactly parallel throughout the postrest trials, which suggests an additive rather than multiplicative relationship between I_R and ${}_sI_R$. There are certain weaknesses and peculiarities in Duncan's study (for example, it could be argued that the 5 minutes' practice was not sufficient to attain the threshold of I_R necessary for the development of ${}_sI_R$, the evidence for which has been presented by Kimble, 1950); but on the whole, it favors Hull's formulation regarding inhibition more than it favors those formulations which involve $I_R \times {}_sI_R$. Another study by Starkweather and Duncan (1954) was essentially the same as the previous experiment except that the massed group was given more prerest practice so that performance on the first postrest trial would be the same for both massed and distributed groups. The rest period was 24 hours. Again, when both groups were given massed practice after the rest, their performance curves were approximately parallel, suggesting that there is no interaction between I_R and ${}_sI_R$. It is possible to argue from some of the evidence in this study, however, that the presence of ${}_sI_R$ was not clearly demonstrated.[4]

Better evidence is presented by Bourne and Archer (1956). Groups trained under massed and distributed practice on the pursuit rotor were given 5 minutes' rest, and then all groups performed under massed conditions. The performance curves *converged* in the postrest period. But the convergence consisted of the performance of the previously distributed group *reducing* to that of the massed group. If the $I_R \times {}_sI_R$ formulation were correct, the result should have been just the opposite, with the previously massed group showing an increase up to the level of the distributed group. The prerest practice was more prolonged in this study than in Duncan's, and it can be argued that there was a sufficient amount of ${}_sI_R$ generated to permit the $I_R \times {}_sI_R$ to show itself. Yet, in another motor learning experiment specially designed to determine if there was an interaction between I_R and ${}_sI_R$, Bowen, Ross, and Andrews (1956) failed to find any evidence of interaction. So while the evidence is not definitive on this point, the preponderance of it does not favor the

[4] It seems fairly certain that the concept of ${}_sI_R$ invoked to explain decremental phenomena in motor learning could not represent the same process as the ${}_sI_R$ involved in experimental extinction.

$I_R \times {}_sI_R$ formulation. The issue, however, does not seem beyond a clear-cut experimental test. For example, in the Jones revision $D \times {}_sI_R$ would always have to be greater than $I_R \times {}_sI_R$, because there can be no performance when D is equal to or less than I_R. If this were true, a person practicing on the pursuit rotor over a long period should finally become unable to perform, since ${}_sI_R$ would continue to grow and inhibit performance. After I_R had dissipated, $D \times {}_sI_R$ would approach or equal $D \times {}_sH_R$, and the subject would be unable to perform the pursuit task. Gleitman, Nachmais, and Neisser (1954) were the first to point out this consequence with respect to Hull's formulation. As far as the writer knows, no one has ever found this kind of "extinction" of the pursuit rotor skill. Subjects have been known to practice the pursuit task day after day for months, long after having reached an asymptote for time on target, yet they show no loss of the skill. Hull's formula, on the other hand, can get around this problem, the arguments of Gleitman et al. (1954) notwithstanding. If ${}_sH_R$ and ${}_sI_R$ both reach an asymptote (Hull, 1951), extinction will have occurred when ${}_sI_R = D \times {}_sH_R$. An increase in D will make it possible for $D \times {}_sH_R$ to be greater than the symptote of ${}_sI_R$, so that extinction need never occur if D remains sufficiently high. Indeed, there are instances (Solomon & Wynne, 1954) of absence of extinction in escape and avoidance training in which the drive is a very strong shock-induced fear reaction.

The unlikely prediction made from Hull's theory by Gleitman et al. (1954) that any response, even though always reinforced, would eventually extinguish if it were repeated often enough was directly tested in experiments by Calvin, Clifford, Clifford, Bolden, and Harvey (1956) and Kendrick (1958). Their studies differ in a few details of experimental procedure. Essentially they ran rats down a long alleyway at the end of which the rats received reinforcement *on every trial.* After some hundreds of trials (spread over many days) all the rats ceased running down the alley; they would not leave the starting box for a specified period of time designated as the criterion for "complete" extinction. Though this outcome lends support to Hull's theory, other interpretations are certainly possible (see Mowrer, 1960, pp. 426–432; Prokasy, 1960). The results of the Calvin et al. and Kendrick experiments may well be due to peculiarities of the experimental procedure. If not, one should expect "extinction with reinforcement" to occur in many other kinds of performance, such as a rat's bar pressing or a pigeon's pecking in a Skinner box, and in many types of repetitious motor tasks.

One experiment is highly relevant to theoretical predictions regarding the effects of drive on motor learning. Wasserman (1951), using a motor learning task (alphabet printing) found that high motivation resulted in performance which was significantly superior to that of low motivation (in both massed and distributed practice groups), the difference becoming progressively greater as practice continued. The Jones revision would predict just the opposite. Since D must always be greater than I_R, $D \times {}_sI_R$ would result in greater performance decrement for the highly motivated group. The motivation in this experiment was controlled by the instructions given to the subjects, one group being task-oriented, the other ego-oriented.

$I_R \times {}_sH_R$

This formulation of an interaction

between reactive inhibition and habit strength implies that the decremental effects on performance caused by the conditions producing I_R (effort and rate of response) will be greater for strong than for weak habits. This is patently incorrect, since it is known that there is a *positive* correlation between number of reinforced responses, of which ${}_SH_R$ is a function, and the number of responses emitted during extinction. The $I_R \times {}_SH_R$ formulation would predict just the opposite, i.e., a *negative* correlation between number of reinforcements and number of responses to extinction. This conclusion is not weakened by the fact (for example, Reid, 1953) that in learning to make a discrimination reversal the animals that have had a greater number of prereversal trials learn the reversal more quickly. This phenomenon may be interpreted in terms of the animal's also overlearning the *act* of making a discrimination (in addition to learning to respond differentially to the S^D and S^Δ), which facilitates the learning of the reversal.

${}_SH_R \times {}_SI_R$

This formulation, derived from Iwahara (1957), is subject to the same criticism just made in the case of $I_R \times {}_SH_R$. It implies that the stronger the habit, the more quickly it should extinguish, which certainly is not true.

$K - \dot{I}_R$

The suggestion of Woodworth and Schlosberg (1954), that total inhibition ($\dot{I}_R = I_R + {}_SI_R$) be subtracted from incentive motivation, K (a function of amount of reinforcement), seems plausible, in that extinction involves the withdrawal of incentive. Within the total Hullian formulation, however, the Woodworth and Schlosberg suggestion meets with the same difficulties pointed out in the two previous cases. Thus:

$${}_S\overline{E}_R = D \times (K - I_R - {}_SI_R) \times {}_SH_R$$

In expanded form:

$$\begin{aligned} {}_S\overline{E}_R = {} & D \times K - D \times I_R - D \times {}_SI_R \\ & \times D \times {}_SH_R \times K \times {}_SH_R \\ & - {}_SH_R \times I_R - {}_SH_R - {}_SI_R \end{aligned}$$

Thus we have again all of the elements that have already been criticized. Spence (1956) has argued, on the basis of experimental findings, that D and K are additive rather than multiplicative as in Hull. But here again the defects of the Woodworth and Schlosberg suggestion of $K - \dot{I}_R$ are evident.

$${}_S\overline{E}_R = (D + K - I_R) \times {}_SH_R$$

Expanded:

$${}_S\overline{E}_R = D \times {}_SH_R + K \times {}_SH_R - {}_SH_R \times \dot{I}_R$$

The last term in the expanded formula again meets with the same difficulty pointed out above. It must be concluded that the $K - \dot{I}_R$ formulation is not an improvement on Hull or Spence.

SUMMARY

Several attempts to reformulate Hull's theory with respect to the inhibition postulates have been criticized. Because of the limitations of both Hull and his revisers in the exact quantification of intervening variables, much of the choice between alternative versions of the theory must be made on the basis of *plausibility* of congruence with empirical findings rather than of *prediction* of these findings in the rigorous sense of the term. All of the attempted revisions to date, with the possible exception of that of Spence, have serious shortcomings in the light of experimental evidence. They cannot, therefore, be regarded as improvements over Hull's original formulation of reaction potential. Advances will be made, not by the mere algebraic manipulation of Hull's intervening variables—the method that characterizes the present attempts—but by the postulation and quantification of new intervening variables, along with the experimental investigation of their interactions.

REFERENCES

Barry, H. Effects of strength of drive on learning and extinction. *J. exp. Psychol.*, 1958, **55**, 473–481.

Bass, M. J., & Hull, C. L. The irradiation of tactile conditioned reflex in man. *J. comp. Psychol.*, 1934, **17**, 47–65.

Bitterman, M. E., & Holtzman, W. H. Conditioning and extinction of the galvanic skin response as a function of anxiety. *J. abnorm. soc. Psychol.*, 1952, **47**, 615–623.

Bourne, L. E., Jr., & Archer, E. J. Time continuously on target as a function of distribution of practice. *J. exp. Psychol.*, 1956, **51**, 25–33.

Bowen, J. H., Ross, S., & Andrews, T. G. A note on the interaction of conditioning and reactive inhibition in pursuit tracking. *J. gen. Psychol.*, 1956, **55**, 153–162.

Brandauer, C. M. A confirmation of Webb's data concerning the action of irrelevant drives. *J. exp. Psychol.*, 1953, **45**, 150–152.

Broadhurst, P. L. Emotionality and the Yerkes-Dodson law. *J. exp. Psychol.*, 1957, **54**, 345–352.

Brown, Janet L. The effect of drive on learning with secondary reinforcement. *J. comp. physiol. Psychol.*, 1956, **49**, 254–260.

Bullock, D. H. The inter-relationship of operant level, extinction ratio, and reserve. *J. exp. Psychol.*, 1950, **40**, 802–804.

Bullock, D. H., & Smith, W. C. An effect of repeated conditioning-extinction upon operant strength. *J. exp. Psychol.*, 1953, **46**, 349–352.

Calvin, A. A., Clifford, T., Clifford, B., Bolden, L., & Harvey, J. An experimental validation of conditioned inhibition. *Psychol. Rep.*, 1956, **2**, 51–56.

Cautela, J. R. Experimental extinction and drive during extinction in a discrimination habit. *J. exp. Psychol.*, 1956, **51**, 299–302.

Cotton, J. W. On making predictions from Hull's theory. *Psychol. Rev.*, 1955, **67**, 303–314.

Crocetti, C. P. The relation of extinction responding to drive level in the white rat. Unpublished doctoral dissertation, Columbia University, 1952.

Dinsmoor, J. A. The effect of hunger on discriminated responding. *J. abnorm. soc. Psychol.*, 1952, **47**, 67–72.

Duncan, C. P. The effect of unequal amounts of practice on motor learning before and after rest. *J. exp. Psychol.*, 1951, **42**, 257–264.

Ellson, D. G. The concept of reflex reserve. *Psychol. Rev.*, 1939, **46**, 566–575.

Eysenck, H. J. "Warm-up" in pursuit rotor learning as a function of the extinction of conditioned inhibition. *Acta Psychol., Amst.*, 1956, **12**, 349–370.

Eysenck, H. J. *The dynamics of anxiety and hysteria.* London: Routledge & Kegan Paul, 1957.

Fitzwater, M. E. The relative effect of reinforcement and nonreinforcement in establishing a form discrimination. *J. comp. physiol. Psychol.*, 1952, **45**, 476–481.

Gleitman, H., Nachmais, J., & Neisser, U. The S-R reinforcement theory of extinction. *Psychol. Rev.*, 1954, **61**, 23–33.

Grice, G. R. The acquisition of a visual discrimination habit following response to a single stimulus. *J. exp. Psychol.*, 1948, **38**, 633–642.

Grice, G. R. Visual discrimination learning with simultaneous and successive presentation of stimuli. *J. comp. physiol. Psychol.*, 1949, **42**, 365–373.

Hanson, H. M. Discrimination training effect on stimulus generalization gradient for spectrum stimuli. *Science*, 1957, **125**, 888–889.

Hilgard, E. R. *Theories of learning.* (2nd ed.) New York: Appleton-Century-Crofts, 1956.

Hilgard, E. R. Intervening variables, hypothetical constructs, parameters, and constants. *Amer. J. Psychol.*, 1958, **71**, 238–246.

Hilgard, E. R., Jones, L. V., & Kaplan, S. J. Conditioned discrimination as related to anxiety. *J. exp. Psychol.*, 1951, **42**, 94–99.

Hilgard, E. R., & Marquis, D. M. *Conditioning and learning.* New York: Appleton-Century-Crofts, 1940.

Hovland, C. I. The generalization of conditioned responses: The sensory generalization of conditioned responses with varying frequencies of tone. *J. gen. Psychol.*, 1937, **17**, 125–148.

Hull, C. L. *Principles of behavior.* New York: Appleton-Century-Crofts, 1943.

Hull, C. L. *Essentials of behavior.* New Haven: Yale Univer. Press, 1951.

Hull, C. L. *A behavior system.* New Haven: Yale Univer. Press, 1952.

Iwahara, S. Hull's concept of inhibition: A revision. *Psychol. Rep.*, 1957, **3**, 9–10.

Jenkins, W. O., & Daugherty, Georgett. Drive and the asymptote of extinction. *J. comp. physiol. Psychol.*, 1951, **44**, 372–377.

Jones, H. G. The status of inhibition in Hull's system: A theoretical revision. *Psychol. Rev.*, 1958, **65**, 179–182.

Kendrick, D. C. Inhibition with reinforcement (conditioned inhibition). *J. exp. Psychol.*, 1958, **56**, 313–318.

Kimble, G. A. Evidence for the role of motivation in determining the amount of reminiscence in pursuit rotor learning. *J. exp. Psychol.*, 1950, **40**, 248–253.

Koch, S. Clark L. Hull. In W. K. Estes, K. MacCorquodale, P. E. Meehl, C. G. Mueller, W. N. Schoenfeld, & W. S. Verplanck (Eds.), *Modern learning theory: A critical analysis of five examples.* New York: Appleton-Century-Crofts, 1954. Pp. 1–176.

Lewis, D. J., & Cotton, J. W. Learning and performance as a function of drive strength during acquisition and extinction. *J. comp. physiol. Psychol.*, 1957, **50**, 189–194.

Liberman, A. M. A comparison of transfer effects during acquisition and extinction of two instrumental responses. *J. exp. Psychol.*, 1951, **41**, 192–198.

MacCorquodale, K., & Meehl, P. E. On a distinction between hypothetical constructs and intervening variables. *Psychol. Rev.*, 1948, **55**, 95–107.

Meyer, D. R. Food deprivation and discrimination reversal learning of monkeys. *J. exp. Psychol.*, 1951, **41**, 10–16.

Miles, R. C. Discrimination in the squirrel monkey as a function of deprivation and problem difficulty. *J. exp. Psychol.*, 1959, **57**, 15–19.

Mowrer, O. H. *Learning theory and behavior.* New York: Wiley, 1960.

Osgood, C. E. *Method and theory in experimental psychology.* New York: Oxford Univer. Press, 1953.

Pavlov, I. P. *Conditioned reflexes.* London: Oxford Univer. Press, 1927.

Perin, C. T. Behavior potentiality as a joint function of the amount of training and degree of hunger at the time of extinction. *J. exp. Psychol.*, 1942, **30**, 93–113.

Perkins, C. C., Jr., & Cacioppo, A. J. The effect of intermittent reinforcement on the change in extinction rate following successive reconditionings. *J. exp. Psychol.*, 1950, **40**, 794–801.

Prokasy, W. F. Postasymptotic performance decrements during massed reinforcements. *Psychol. Bull.*, 1960, **57**, 237–247.

Razran, G. Extinction re-examined and reanalyzed: A new theory. *Psychol. Rev.*, 1956, **63**, 39–52.

Razran, G. H. S. Transposition of relational responses and generalization of conditioned responses. *Psychol. Rev.*, 1938, **45**, 532–538.

Reid, L. S. The development of noncontinuity behavior through continuity learning. *J. exp. Psychol.*, 1953, **46**, 107–112.

Reynolds, B. The acquisition of a trace conditioned response as a function of the magnitude of the stimulus trace. *J. exp. Psychol.*, 1945, **35**, 15–30. (a)

Reynolds, B. Extinction of trace conditioned responses as a function of the spacing of trials during the asquisition and extinction series. *J. exp. Psychol.*, 1945, **35**, 81–95. (b)

Reynolds, B., Marx, M. H., & Henderson, R. L. Resistance to extinction as a function of drive-reward interaction. *J. comp. physiol. Psychol.*, 1952, **45**, 36–42.

Sackett, R. S. The effect of strength of drive at the time of extinction upon resistance to extinction in rats. *J. comp. Psychol.*, 1939, **27**, 411–431.

Saltzman, I., & Koch, S. The effect of low intensities of hunger on the behavior mediated by a habit of maximum strength. *J. exp. Psychol.*, 1948, **38**, 347–370.

Sheffield, Virginia F. Resistance to extinction as a function of the distribution of extinction trials. *J. exp. Psychol.*, 1950, **40**, 305–313.

Shurrager, P. S., & Shurrager, H. C. Rate of learning measured at a single synapse. *J. exp. Psychol.*, 1946, **36**, 347–354.

Skinner, B. F. *The behavior of organisms.* New York: Appleton-Century-Crofts, 1938.

Smith, M. H., & Hay, W. J. Rate of response during operant discriminations. *J. exp. Psychol.*, 1954, **48**, 259–264.

Solomon, R. L., & Wynne, L. C. Traumatic avoidance learning: The principles of anxiety conservation and partial irreversibility. *Psychol. Rev.*, 1954, **61**, 353–385.

Spence, K. W. The differential response in animals to stimuli varying within a single dimension. *Psychol. Rev.*, 1937, **44**, 430–444.

Spence, K. W. *Behavior theory and conditioning.* New Haven: Yale Univer. Press, 1956.

Spence, K. W., & Farber, I. E. Conditioning and extinction as a function of anxiety. *J. exp. Psychol.*, 1953, **45**, 116–119.

Spence, K. W., & Farber, I. E. The relation of anxiety to differential eyelid conditioning. *J. exp. Psychol.*, 1954, **47**, 127–134.

Spence, K. W., Goodrich, K. P., & Ross, L. E. Performance in differential conditioning and discrimination learning as a function of hunger and relative response frequency. *J. exp. Psychol.*, 1959, **58**, 8–16.

Stanley, W. C. Extinction as a function of the spacing of extinction trials. *J. exp. Psychol.*, 1952, **43**, 246–260.

Starkweather, J. A., & Duncan, C. P. A test for conditioned inhibition in motor learning. *J. exp. Psychol.*, 1954, **47**, 351–356.

Teel, K. S. Habit strength as a function of motivation during learning. *J. comp. physiol. Psychol.*, 1952, **45**, 188–191.

Wasserman, H. N. The effect of motivation and amount of pre-rest practice upon inhibitory potential in motor learning. *J. exp. Psychol.*, 1951, **42**, 162–172.

Webb, W. B. A test of "relational" vs. "specific stimulus" learning in discrimination problems. *J. comp. physiol. Psychol.*, 1950, **43**, 70–72.

Woodworth, R. S., & Schlosberg, H. *Experimental psychology.* (Rev. ed.) New York: Holt, 1954.

Yerkes, R. M., & Dodson, J. D. The relation of strength of stimulus to rapidity of habit formation. *J. comp. Neurol.*, 1908, **18**, 459–482.

(Received February 11, 1960)

Psychological Bulletin
1961, Vol. 58, No. 4, 299–304

ACQUIESCENCE AND THE FACTORIAL INTERPRETATION OF THE MMPI[1]

SAMUEL MESSICK
Educational Testing Service

AND

DOUGLAS N. JACKSON
Pennsylvania State University

The operation of reliable response sets or stylistic consistencies has been frequently noted on personality and attitude scales with a true-false or agree-disagree format (cf. Cronbach, 1946, 1950; Fricke, 1956; Messick & Jackson, 1958). It has recently been conjectured (Jackson & Messick, 1958) that the major common factors in personality inventories of this type are interpretable primarily in terms of such stylistic consistencies rather than in terms of specific item content. The present paper attempts to annotate the influence of two response styles, the tendency to agree or acquiesce and the tendency to respond in a desirable way, using the Minnesota Multiphasic Personality Inventory (MMPI) as an example of inventories with this general response form. In particular, a high correlation will be noted between factor loadings on the largest factor, as obtained in several published factor analyses of the MMPI, and certain indices of acquiescence.

Barnes (1956b), in evaluating the Berg (1955) deviation hypothesis on the MMPI, found that the tendency to answer atypically or deviantly "true" was highly correlated with scores on the psychotic scales, and the tendency to answer atypically "false" was highly correlated with the neurotic triad. This result is consistent with the fact, noted by Cottle and Powell (1951) and others (Barnes, 1956b; Fricke, 1956), that a large proportion of MMPI psychotic items are keyed true and a large proportion of neurotic items keyed false, suggesting that differential tendencies to respond atypically "true" and "false" might have been involved in the discrimination of criterion groups upon which the scoring keys were based. Barnes (1956a) also pointed out a marked similarity between the correlations of MMPI scales with these two deviant response tendencies and factor loadings for the scales on the two major factors reported by Wheeler, Little, and Lehner (1951); he concluded that the number of atypical true answers is a "pure factor test" of the first or "psychotic" factor and that the number of deviant false answers has a high loading on the second or "neurotic" factor. The two major MMPI factors obtained by Welsh (1956) also displayed a similar pattern of loadings, and it is noteworthy that the "pure factor" reference scale *A* which Welsh developed for his first or "anxiety" factor had 38 out of 39 items keyed true, while the reference scale *R* for the second or "repression" factor had all 40 of its items keyed false.

In view of the striking similarity between the effects of consistent tendencies to respond "true" and "false" and patterns of factor loadings obtained in two studies of

[1] This study is part of a larger project on stylistic determinants in clinical personality assessment supported by the National Institute of Mental Health, United States Public Health Service, under Research Grants M-2878 to Educational Testing Service and M-2738 to Pennsylvania State University. The authors wish to thank George S. Welsh for graciously supplying scoring keys for his "pure" MMPI scales and Philip E. Slater for making available his factor analyses of the MMPI.

MMPI scales, all factor analyses of the MMPI readily available in the literature were reviewed, in order to evaluate the possible relationship between each scale's factor loading on the major factor and an index of its potential for reflecting acquiescence. The particular index of acquiescence used was the proportion of items keyed true on each scale, which, assuming that the acquiescence-evoking properties of items are uniform over all MMPI scales, can be considered to reflect the extent to which total scores on a scale are influenced by consistent tendencies to respond "true." High scores on a scale with a large proportion of items keyed true would thus be assumed to reflect a general tendency to acquiesce, in addition, of course, to the contribution of other stylistic tendencies and of systematic content responses. Jackson (1960) used this index to evaluate the effects of acquiescence on the California Psychological Inventory, and Voas (1958) used the proportion of items keyed false as a criterion for constructing response bias scales. Voas (1958) also estimated loadings for scales from the MMPI and the Guilford-Zimmerman Temperament Survey on a factor marked by two measures of the tendency to respond "false" and found that these loadings correlated .86 with the proportion of items keyed false on each scale. These findings support the use of the index in the present context.

Factor loadings for MMPI scales were obtained from eight studies by Abrams (1949, summarized by French, 1953), Cook and Wherry (1950), Cottle (1950), Tyler (1951), Wheeler, Little, and Lehner (1951), Welsh (1956), Slater (1958), and Kassebaum, Couch, and Slater (1959). A fairly uniform finding from these studies is that only two major factors and two or three minor ones are necessary to account for interrelations among the scales. Spearman rank correlations were computed between loadings on the largest factor in each study and the proportion of items keyed true on each scale; the results are summarized in Table 1. In some of the factor analyses, values were not reported for scales with small loadings on the factor, so in computing correlation coefficients these scales were considered to be tied at an appropriate rank below scales with reported positive loadings and above scales with reported negative loadings. Corrections for ties (cf. Siegel, 1956) were computed for two of the studies with the most scales tied at the same rank (Wheeler, Little, & Lehner's normal sample and Tyler's sample), but the coefficients changed only .01.

Of 11 different subject samples represented in these eight studies, significant correlations were obtained for 8 of them, four of the coefficients exceeding .85. These strikingly consistent findings indicate that in most of these studies the largest factor on the MMPI is interpretable in terms of acquiescence. In evaluating the few apparently inconsistent results, it is important to note that for Abrams's (1949) neurotic sample, the correlation with the largest factor was −.15, but with the second largest it was .52. Also, in Tyler's (1951) study the correlation with the largest rotated factor was .33, but with the unrotated first centroid it was .52, $p<.05$. These findings suggest that for those studies in which the correspondence between the proportion of items keyed true and the factor loadings was not close, the factor structures could have been rotated to produce a higher correlation. Ana-

TABLE 1

Spearman Rank Correlation (ρ) between Factor Loadings on the Largest MMPI Factor and Proportion of Items Keyed "True" on Each Scale

Study	Scales Included	Sample	ρ
Abrams, 1949	11 scales: *L, F, Hs, D, Hy, Pd, Mf, Pa, Pt, Sc, Ma*	117 normal male veterans	.907**
		201 neurotic male veterans	−.148 (largest factor) .516 (2nd largest)
Cook & Wherry, 1950	11 scales: *L, F, Hs, D, Hy, Pd, Mf, Pa, Pt, Sc, Ma*	111 male naval submarine candidates	.605*
Cottle, 1950	11 scales: *L, F, Hs, D, Hy, Pd, Mf, Pa, Pt, Sc, Ma*	400 male veterans	.916**
Tyler, 1951	15 scales: *Hs, D, Hy, Pd, Mf, Pa, Pt, Sc, Ma, Si, St, Pr, Ac, Re, Do*	107 female graduate students	.328
Wheeler, Little, & Lehner, 1951	12 scales: *L, K, F, Hs, D, Hy, Pd, Mf, Pa, Pt, Sc, Ma*	112 male college students	.558
		110 male neuropsychiatric patients	.874**
Welsh, 1956	11 pure scales: *K′, Hs′, D′, Hy′, Pd′, Mf′, Pa′, Pt′, Sc′, Ma′, Si′*	150 male VA general hospital patients	.870**
	11 pure scales plus *A, Gm, Ja, R*	Same 150 males	.897**
Slater, 1958	43 scales: *L, F, K, Hs, D, Hy, Pd, Mf, Pa, Pt, Sc, Ma, Si, Nm, Dp, Fm, A, R, Im, Pr, To, C, P, Sp, Rp, Sy, Re, St, Lp, Do, Es, Ie, Ac, Ai, O–I, Lb, Ne, Ca, Pl, Ht, Cht,* Z_1, Z_2	102 aged males	.728**
		109 aged females	.718**
Kassebaum, Couch, & Slater, 1959	32 scales: *L, F, K, Hs, D, Hy, Pd, Mf, Pa, Pt, Sc, Ma, Si, Es, Ie, Lp, Ai, Sy, Ac, Re, Do, Pr, St, Im, Sp, Fm, Rp, R, A, Dp, To, OI*	160 Harvard College freshmen	.625**

* p <.05.
** p <.01.

lytical procedures similar to the computation of B weights in multiple correlation analysis are available (Mosier, 1939) for rotating to maximize the correlation between a factor and a criterion, which in this case would be a vector of proportions of true items. However, an adequate application of this technique requires loadings for all the scales on the factors under consideration, and for those studies providing this information (e.g., Welsh, 1956) there was usually little need to rotate.

Another consideration which suggests that a rotation of axes might clarify the role of acquiescence on the MMPI is the fact that scales with high loadings on the second largest MMPI factor usually tend to have a high proportion of false items in their keys. Kassebaum, Couch, and Slater (1959) noticed this in their factor results and suggested that their second factor partly reflected a general tendency to respond "false." Although correlations between the proportion of items keyed true and loadings on the second MMPI factor are usually not nearly as high as correlations with the first factor, some significant coefficients occur; e.g., the correlation between the proportion of items keyed true and loadings on the second factor in the study by Kassebaum, Couch, and Slater (1959) was −.44, p<.05 with 30 *df*, and in Welsh's (1956) study it was −.64, p<.05 with 13 *df*.

This result is consistent with Barnes' (1956a) finding of a correspondence between atypical true answers and the first MMPI factor and atypical false answers and the second factor. Since these two factors are usually orthogonal, this correspondence might be considered evidence for two relatively independent response biases, one a tendency to agree and the other to disagree. Such a contention is consistent with Barnes' (1956b) finding of a correlation of .11 between deviant responses answered "true" and "false" and with the fact that Welsh's (1956) *A* and *R* scales are usually only slightly negatively correlated. Although these results cannot be accounted for by a simple response set of acquiescence, it is not necessary to postulate two independent sets to agree and to disagree. As has been pointed out (Jackson & Messick, 1958), all that is required to account for the findings is the operation of at least one other factor in conjunction with acquiescence. Thus, the *A* scale can have a high positive loading on an acquiescence factor and the *R* scale a high negative loading, yet the two scales could be uncorrelated if they both had positive, or negative, loadings on some other dimension. Other factors which could moderate the operation of acquiescence on the MMPI might be specific content dimensions or some other response style. As previously suggested (Jackson & Messick, 1958), a particularly likely candidate for such a role is the stylistic tendency to respond in a desirable way.

Possible influences on MMPI scores of a set to respond desirably have been widely documented (cf. De Soto & Kuethe, 1959; Edwards, 1957; Fordyce, 1956; Hanley, 1956, 1957; Jackson & Messick, 1958; Taylor, 1959; Wiggins & Rumrill, 1959). Fordyce (1956), for example, has noted a marked similarity between loadings on the largest MMPI factor from Wheeler, Little, and Lehner's (1951) psychiatric sample and correlations of MMPI scales with a measure of desirability. In fact, the rank correlation between the loadings and the correlation coefficients is approximately −.75, and since the proportion of items keyed true on each MMPI scale correlates only about −.50 with the desirability coefficients, it seems likely that a combination of desirability and acquiescence would lead to even better prediction of the factor (cf. Messick, 1959). Although this and some other reported relationships are somewhat equivocal because the measures of desirability used were partially confounded with acquiescence, e.g., Edwards' *SD* scale and Hanley's *Ex* scale, high correlations have also been reported between MMPI scales and desirability measures having a balanced number of true and false items (Edwards, 1957; Hanley, 1957; Wiggins & Rumrill, 1959).

In an attempt to take these findings into account, it is suggested that the acquiescence-evoking properties of items are not, as assumed above, uniform over all scales, but that acquiescence is elicited differentially as a function, perhaps, of specific item content, of the clarity or ambiguity with which the content is stated, and in particular of the perceived desirability of the statement. In the extreme, it is suggested that the two major factors usually found for the MMPI may be rotated into positions interpretable as two response styles—the tendency to acquiesce and the tendency to respond desirably. The negative poles of these dimensions would be the tendencies to disagree and to respond undesirably, respectively. Response variance on MMPI scales would then be primarily a function of these two stylistic components in various weighted proportions. Studies including independent marker variables for the two styles are of course required to identify the factor positions. Much research is also needed into the precise nature of the set to respond desirably, particularly in view of three complicating results: (*a*) the finding of consistent individual differences in judgments of desirability (Messick, 1960); (*b*) the distinction between personal and social desirability (Borislow, 1958; Rosen, 1956); and (*c*) the differentiation between a tendency to endorse certain desirable items which exhibit large mean shifts under desirability instructions and the tendency to endorse other desirble items which presumably reflect a group norm (Voas, 1958; Wiggins, 1959).

In conclusion, the findings offer clear evidence that acquiescence, as moderated by item desirability, plays a dominant role in personality inventories like the MMPI. Focused empirical investigations are required to develop a refined interpretation of these and other stylistic consistencies in terms of personality organization and psychopathology.

REFERENCES

Abrams, E. N. A comparative factor analytic study of normal and neurotic veterans. Unpublished doctoral dissertation, University of Michigan, 1949.

Barnes, E. H. Factors, response bias, and the MMPI. *J. consult. Psychol.*, 1956, **20**, 419–421. (a)

Barnes, E. H. Response bias and the MMPI. *J. consult. Psychol.*, 1956, **20**, 371–374. (b)

Berg, I. A. Response bias and personality: The deviation hypothesis. *J. Psychol.*, 1955, **40**, 61–72.

Borislow, B. The Edwards Personal Preference Schedule (EPPS) and fakability. *J. appl. Psychol.*, 1958, **42**, 22–27.

Cook, E. B., & Wherry, R. J. A factor analysis of MMPI and aptitude test data. *J. appl. Psychol.*, 1950, **34**, 260–266.

Cottle, W. C. A factorial study of the Multiphasic, Strong, Kuder, and Bell inventories using a population of adult males. *Psychometrika*, 1950, **15**, 25–47.

Cottle, W. C., & Powell, J. O. The effect of random answers to the MMPI. *Educ. psychol. Measmt*, 1951, **11**, 224–227.

Cronbach, L. J. Response sets and test validity. *Educ. psychol. Measmt*, 1946, **6**, 475–494.

Cronbach, L. J. Further evidence on response sets and test design. *Educ. psychol. Measmt*, 1950, **10**, 3–31.

De Soto, C. B., & Kuethe, J. L. The set to claim undesirable symptoms in personality inventories. *J. consult. Psychol.*, 1959, **23**, 496–500.

Edwards, A. L. *The social desirability variable in personality assessment and research.* New York: Dryden, 1957.

Fordyce, W. E. Social desirability in the MMPI. *J. consult. Psychol.*, 1956, **20**, 171–175.

French, J. W. *The description of personality measurements in terms of rotated factors.* Princeton, N. J.: Educational Testing Service, 1953.

Fricke, B. G. Response set as a suppressor variable in the OAIS and MMPI. *J. consult. Psychol.*, 1956, **20**, 161–169.

Hanley, C. Social desirability and responses to items from three MMPI scales: *D*, *Sc*, and *K*. *J. appl. Psychol.*, 1956, **40**, 324–328.

Hanley, C. Deriving a measure of test-taking defensiveness. *J. consult. Psychol.*, 1957, **21**, 391–397.

Jackson, D. N. Stylistic response determinants in the California Psychological Inventory. *Educ. psychol. Measmt*, 1960, **20**, 339–346.

Jackson, D. N., & Messick, S. Content and style in personality assessment. *Psychol. Bull.*, 1958, **55**, 243–252.

Kassebaum, G. G., Couch, A. S., & Slater, P. E. The factorial dimensions of the MMPI. *J. consult. Psychol.*, 1959, **23**, 226–236.

Messick, S. Review of Allen Edwards', *The social desirability variable in personality assessment and research.* *Educ. psychol. Measmt.*, 1959, **19**, 451–454.

Messick, S. Dimensions of social desirability.

J. consult. Psychol., 1960, **24**, 279–287.

Messick, S., & Jackson, D. N. The measurement of authoritarian attitudes. *Educ. psychol. Measmt*, 1958, **18**, 241–253.

Mosier, C. I. Determining a simple structure when loadings for certain tests are known. *Psychometrika*, 1939, **4**, 149–162.

Rosen, E. Self-appraisal, personal desirability, and perceived social desirability of personality traits. *J. abnorm. soc. Psychol.*, 1956, **52**, 151–158.

Siegel, S. *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956.

Slater, P. E. Personality structure in old age. Progress Report, 1958, Age Center of New England, Project M-1402, National Institute of Mental Health.

Taylor, J. B. Social desirability and MMPI performance: The individual case. *J. consult. Psychol.*, 1959, **23**, 514–517.

Tyler, F. T. A factorial analysis of fifteen MMPI scales. *J. consult. Psychol.*, 1951, **15**, 451–456.

Voas, R. B. Relationships among three types of response sets. Report No. 15, 1958, Naval School of Aviation Medicine, Pensacola, Project NM 16 0111 Subtask 1.

Welsh, G. S. Factor dimensions *A* and *R*. In G. S. Welsh & W. G. Dahlstrom (Eds.), *Basic readings on the MMPI in psychology and medicine.* Minneapolis: Univer. Minnesota Press, 1956.

Wheeler, W. M., Little, K. B., & Lehner, G. F. J. The internal structure of the MMPI. *J. consult. Psychol.*, 1951, **15**, 134–141.

Wiggins, J. S. Interrelationships among MMPI measures of dissimulation under standard and social desirability instructions. *J. consult. Psychol.*, 1959, **23**, 419–427.

Wiggins, J. S., & Rumrill, C. Social desirability in the MMPI and Welsh's factor scales *A* and *R*. *J. consult. Psychol.*, 1959, **23**, 100–106.

(Received April 7, 1960)

Psychological Bulletin
1961, Vol. 58, No. 4, 305-316

SCALES AND STATISTICS: PARAMETRIC AND NONPARAMETRIC[1]

NORMAN H. ANDERSON
University of California, Los Angeles

The recent rise of interest in the use of nonparametric tests stems from two main sources. One is the concern about the use of parametric tests when the underlying assumptions are not met. The other is the problem of whether or not the measurement scale is suitable for application of parametric procedures. On both counts parametric tests are generally more in danger than nonparametric tests. Because of this, and because of a natural enthusiasm for a new technique, there has been a sometimes uncritical acceptance of nonparametric procedures. By now a certain degree of agreement concerning the more practical aspects involved in the choice of tests appears to have been reached. However, the measurement theoretical issue has been less clearly resolved. The principal purpose of this article is to discuss this latter issue further. For the sake of completeness, a brief overview of practical statistical considerations will also be included.

A few preliminary comments are needed in order to circumscribe the subsequent discussion. In the first place, it is assumed throughout that the data at hand arise from some sort of measuring scale which gives numerical results. This restriction is implicit in the proposal to compare parametric and nonparametric tests since the former do not apply to strictly categorical data (but see Cochran, 1954). Second, parametric tests will mean tests of significance which assume equinormality, i.e., normality and some form of homogeneity of variance. For convenience, parametric test, F test, and analysis of variance will be used synonymously. Although this usage is not strictly correct, it should be noted that the t test and regression analysis may be considered as special applications of F. Nonparametric tests will refer to significance tests which make considerably weaker distributional assumptions as exemplified by rank order tests such as the Wilcoxon T, the Kruskal-Wallis H, and by the various median-type tests. Third, the main focus of the article is on tests of significance with a lesser emphasis on descriptive statistics. Problems of estimation are touched on only slightly although such problems are becoming increasingly important.

Finally, a word of caution is in order. It will be concluded that parametric procedures constitute the everyday tools of psychological statistics, but it should be realized that any area of investigation has its own statistical peculiarities and that general statements must always be adapted to the prevailing practical situation. In many cases, as in pilot work, for instance, or in situations in which data are cheap and plentiful, nonparametric tests, shortcut parametric tests (Tate & Clelland, 1957), or tests by visual inspection may well be the most efficient.

[1] An earlier version of this paper was presented at the April 1959 meetings of the Western Psychological Association. The author's thanks are due F. N. Jones and J. B. Sidowski for their helpful comments.

Practical Statistical Considerations

The three main points of comparison between parametric and nonparametric tests are significance level, power, and versatility. Most of the relevant considerations have been treated adequately by others and only a brief summary will be given here. For more detailed discussion, the articles of Cochran (1947), Savage (1957), Sawrey (1958), Gaito (1959), and Boneau (1960) are especially recommended.

Significance level. The effects of lack of equinormality on the significance level of parametric tests have received considerable study. The two handiest sources for the psychologist are Lindquist's (1953) citation of Norton's work, and the recent article of Boneau (1960) which summarizes much of the earlier work. The main conclusion of the various investigators is that lack of equinormality has remarkably little effect although two exceptions are noted: one-tailed tests and tests with considerably disparate cell *n*'s may be rather severely affected by unequal variances.[2]

A somewhat different source of perturbation of significance level should also be mentioned. An overall test of several conditions may show that something is significant but will not localize the effects. As is well known, the common practice of *t* testing pairs of means tends to inflate the significance level even when the over-all *F* is significant. An analogous inflation occurs with nonparametric tests. There are parametric multiple comparison procedures which are rigorously applicable in many such situations (Duncan, 1955; Federer, 1955) but analogous nonparametric techniques have as yet been developed in only a few cases.

Power. As Dixon and Massey (1957) note, rank order tests are nearly as powerful as parametric tests under equinormality. Consequently, there would seem to be no pressing reason in most investigations to use parametric techniques for reasons of power *if* an appropriate rank order test is available (but see Snedecor, 1956, p. 120). Of course, the loss of power involved in dichotomizing the data for a median-type test is considerable.

Although it might thus be argued that rank order tests should be generally used where applicable, it is to be suspected that such a practice would produce negative transfer to the use of the more incisive experimental designs which need parametric analyses. The logic and computing rules for the analysis of variance, however, follow a uniform pattern in all situations and thus provide maximal positive transfer from the simple to the more complex experiments.

There is also another aspect of power which needs mention. Not infrequently, it is possible to use existing data to get a rough idea of the chances of success in a further related experiment, or to estimate the *N* required for a given desired probability of success (Dixon & Massey, 1957, Ch. 14). Routine methods are available for these purposes when parametric statistics are employed but similar procedures are available only for certain nonparametric tests such as chi square.

[2] The split-plot designs (e.g., Lindquist, 1953) commonly used for the analysis of repeated or correlated observations have been subject to some criticism (Cotton, 1959; Greenhouse & Geisser, 1959) because of the additional assumption of equal correlation which is made. However, tests are available which do not require this assumption (Cotton, 1959; Greenhouse & Geisser, 1959; Rao, 1952).

Versatility. One of the most remarkable features of the analysis of variance is the breadth of its applicability, a point which has been emphasized by Gaito (1959). For present purposes, the ordinary factorial design will serve to exemplify the issue. Although factorial designs are widely employed, their uses in the investigation and control of minor variables have not been fully exploited. Thus, Feldt (1958) has noted the general superiority of the factorial design in matching or equating groups, an important problem which is but poorly handled in current research (Anderson, 1959). Similarly, the use of replications as a factor in the design makes it possible to test and partially control for drift or shift in apparatus, procedure, or subject population during the course of an experiment. In the same way, taking experimenters or stimulus materials as a factor allows tests which bear on the adequacy of standardization of the experimental procedures and on the generalizability of the results.

An analogous argument could be given for latin squares, largely rehabilitated by the work of Wilk and Kempthorne (1955), which are useful when subjects are given successive treatments; for orthogonal polynomials and trend tests for correlated scores (Grant, 1956) which give the most sensitive tests when the independent variable is scaled; as well as for the multivariate analysis of variance (Rao, 1952) which is applicable to correlated dependent variables measured on incommensurable scales.

The point to these examples and to the more extensive treatment by Gaito is straightforward. Their analysis is more or less routine when parametric procedures are used. However, they are handled inadequately or not at all by current nonparametric methods.

It thus seems fair to conclude that parametric tests constitute the standard tools of psychological statistics. In respect of significance level and power, one might claim a fairly even match. However, the versatility of parametric procedures is quite unmatched and this is decisive. Unless and until nonparametric tests are developed to the point where they meet the routine needs of the researcher as exemplified by the above designs, they cannot realistically be considered as competitors to parametric tests. Until that day, nonparametric tests may best be considered as useful minor techniques in the analysis of numerical data.

Too promiscuous a use of F is, of course, not to be condoned since there will be many situations in which the data are distributed quite wildly. Although there is no easy rule with which to draw the line, a frame of reference can be developed by studying the results of Norton (Linquist, 1953) and of Boneau (1960). It is also quite instructive to compare p values for parametric and nonparametric tests of the same data.

It may be worth noting that one of the reasons for the popularity of nonparametric tests is probably the current obsession with questions of statistical significance to the neglect of the often more important questions of design and power. Certainly some minimal degree of reliability is generally a necessary justification for asking others to spend time in assessing the importance of one's data. However, the question of statistical significance is only a first step, and a relatively minor one at that, in the over-all process of evaluating a set of results. To say that a result is statistically significant simply gives reasonable ground for believing that some nonchance effect was obtained. The meaning of a nonchance effect rests on an assessment of the design of the investigation. Even with judicious design, however, phenomena are seldom pinned down in a single study so that the question of replicability in further work often arises also. The statistical aspects of these these two questions are not without importance but tend to be neglected when too heavy an emphasis is placed on p values. As has been noted, it is the parametric procedures which are the more useful in both respects.

Measurement Scale Considerations

The second and principal part of the article is concerned with the relations between types of measurement scales and statistical tests. For convenience, therefore, it will be assumed that lack of equinormality presents no serious problem. Since the F ratio remains constant with changes in unit or zero point of the measuring scale, we may ignore ratio scales and consider only ordinal and interval scales. These scales are defined following Stevens (1951). Briefly, an ordinal scale is one in which the events measured are, in some empirical sense, ordered in the same way as the arithmetic order of the numbers assigned to them. An interval scale has, in addition, an equality of unit over different parts of the scale. Stevens goes on to characterize scale types in terms of permissible transformations. For an ordinal scale, the permissible transformations are monotone since they leave rank order unchanged. For an interval scale, only the linear transformations are permissible since only these leave relative distance unchanged. Some workers (e.g., Coombs, 1952) have considered various scales which lie between the ordinal and interval scales. However, it will not be necessary to take this further refinement of the scale typology into account here.

As before, we suppose that we have a measuring scale which assigns numbers to events of a certain class. It is assumed that this measuring scale is an ordinal scale but not necessarily an interval scale. In order to fix ideas, consider the following example. Suppose that we are interested in studying attitude toward the church. Subjects are randomly assigned to two groups, one of which, reads Communication A, while the other reads Communication B. The subjects' attitudes towards the church are then measured by asking them to check a seven category pro-con rating scale. Our problem is whether the data give adequate reason to conclude that the two communications had different effects.

To ascertain whether the communications had different effects, some statistical test must be applied. In some cases, to be sure, the effects may be so strong that the test can be made by inspection. In most cases, however, some more objective method is necessary. An obvious procedure would be to assign the numbers 1 to 7, say, to the rating scale categories and apply the F test, at least if the data presented some semblance of equinormality. However, some writers on statistics (e.g., Siegel, 1956; Senders, 1958) would object to this on the ground that the rating scale is only an ordinal scale, the data are therefore not "truly numerical," and hence that the operations of addition and multiplication which are used in computing F cannot meaningfully be applied to the scores. There are three different questions involved in this objection, and much of the controversy over scales and statistics has arisen from a failure to keep them separate. Accordingly, these three questions will be taken up in turn.

Question 1. Can the F *test be applied to data from an ordinal scale?* It is convenient to consider two cases of this question according as the assumption of equinormality is satisfied or not. Suppose first that equinormality obtains. The caveat against parametric statistics has been stated most explicitly by Siegel (1956) who says:

> The conditions which must be satisfied . . . before any confidence can be placed in any probability statement obtained by the use of the *t* test are at least these: . . . 4. The variables involved must have been measured in *at least* an interval scale . . . (p. 19). (By permission, from *Nonparametric Statistics*, by S. Siegel. Copyright, 1956. McGraw-Hill Book Company, Inc.)

This statement of Siegel's is completely incorrect. This particular question admits of no doubt whatsoever. The *F* (or *t*) test may be applied without qualm. It will then answer the question which it was designed to answer: can we reasonably conclude that the difference between the means of the two groups is real rather than due to chance? The justification for using *F* is purely statistical and quite straightforward; there is no need to waste space on it here. The reader who has doubts on the matter should postpone them to the discussion of the two subsequent questions, or read the elegant and entertaining article by Lord (1953). As Lord points out, the statistical test can hardly be cognizant of the empirical meaning of the numbers with which it deals. Consequently, the validity of a statistical inference cannot depend on the type of measuring scale used.

The case in which equinormality does not hold remains to be considered. We may still use *F*, of course, and as has been seen in the first part, we would still have about the same significance level in most cases. The *F* test might have less power than a rank order test so that the latter might be preferable in this simple two group experiment. However, insofar as we wish to inquire into the reliability of the difference between the measured behavior of the two groups in our particular experiment, the choice of statistical test would be governed by purely statistical considerations and have nothing to do with scale type.

Question 2. Will statistical results be invariant under change of scale? The problem of invariance of result stems from the work of Stevens (1951) who observes that a statistic computed on data from a given scale will be invariant when the scale is changed according to any given permissible transformation. It is important to be precise about this usage of invariance. It means that if a statistic is computed from a set of scale values and this statistic is then transformed, the identical result will be obtained as when the separate scale values are transformed and the statistic is computed from these transformed scale values.

Now our scale of attitude toward the church is admittedly only an ordinal scale. Consequently, we would expect it to change in the direction of an interval scale in future work. Any such scale change would correspond to a monotone transformation of our original scale since only such transformations are permissible with an ordinal scale. Suppose then that a monotone transformation of the scale has been made subsequent to the experiment on attitude change.

We would then have two sets of data: the responses as measured on the original scale used in the experiment, and the transformed values of these responses as measured on the new, transformed scale. (Presumably, these transformed scale values would be the same as the subjects would have made had the new scale been used in the original experiment, although this will no doubt depend on the experimental basis of the new scale.) The question at issue then becomes whether the same significance results will be obtained from the two sets of data. If rank order tests are used, the same significance results will be found in either case because any permissible transformation leaves rank order unchanged. However, if parametric tests are employed, then different significance statements may be obtained from the two sets of data. It is possible to get a significant *F* from the original data and not from the transformed data, and vice versa. Worse yet, it is even logically possible that the means of the two groups will lie in reverse order on the two scales.

The state of affairs just described is clearly undesirable. If taken uncritically, it would constitute a strong argument for using only rank order tests on ordinal scale data and restricting the use of *F* to data obtained from interval scales. It is the purpose of this section to show that this conclusion is unwarranted. The basis of the argument is that the naming of the scales has begged the psychological question.

Consider interval scales first, and imagine that two students, P and Q, in an elementary lab course are assigned to investigate some process. This process might be a ball rolling on a plane, a rat running an alley, or a child doing sums. The students cooperate in the experimental work, making the same observations, except that they use different measuring scales. P decides to measure time intervals. He reasons that it makes sense to speak of one time interval as being twice another, that time intervals therefore form a ratio scale, and hence a fortiori an interval scale. Q decides to measure the speed of the process (feet per second, problems per minute). By the same reasoning as used by P, Q concludes that he has an interval scale also. Both P and Q are aware of current strictures about

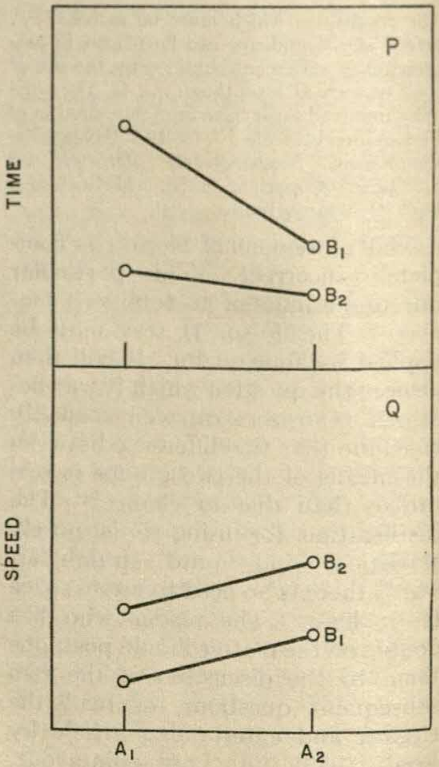


FIG. 1. Temporal aspects of some process obtained from a 2×2 design. (The data are plotted as a function of Variable A with Variable B as a parameter. Subscripts denote the two levels of each variable. Note that Panel P shows an interaction, but that Panel Q does not.)

scales and statistics. However, since each believes (and rightly so) that he has an interval scale, each uses means and applies parametric tests in writing his lab report. Nevertheless, when they compare their reports they find considerable difference in their descriptive statistics and graphs (Figure 1), and in their *F* ratios as well. Consultation with a statistican shows that these differences are direct consequences of the difference in the measuring scales. Evidently then, possession of an interval scale does not guarantee invariance of interval scale statistics.

For ordinal scales, we would expect to obtain invariance of result by using ordinal scale statistics such as the median (Stevens, 1951). Let us suppose that some future investigator finds that attitude toward the church is multidimensional in nature and has, in fact, obtained interval scales for each of the dimensions. In some of his work he chanced to use our original ordinal scale so that he was able to find the relation between this ordinal scale and the multidimensional representation of the attitude. His results are shown in Figure 2. Our ordinal scale is represented by the curved line in the plane of the two dimensions. Thus, a greater distance from the origin as measured along the line stands for a higher value on our ordinal scale. Points A and B on the curve represent the medians of Groups A and B in our experiment, and it is seen that Group A is more pro-church than Group B on our ordinal scale. The median scores for these two groups on the two dimensions are obtained simply by projecting Points A and B onto the two dimensions. All is well on Dimension 2 since there Group A is greater than Group B. On Dimension 1, however, a reversal is found: Group A is less than Group B, contrary to our ordinal scale results. Evidently then, possession of an ordinal scale does not guarantee invariance of ordinal scale statistics.

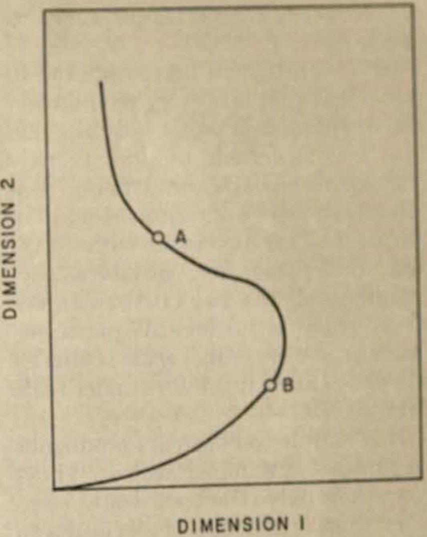


FIG. 2. The curved line represents the ordinal scale of attitude toward the church plotted in the two-dimensional space underlying the attitude. (Points A and B denote the medians of two experimental groups. The graph is hypothetical, of course.)

A rather more drastic loss of invariance would occur if the ordinal scale were measuring the resultant effect of two or more underlying processes. This could happen, for instance, in the study of approach-avoidance conflict, or ambivalent behavior, as might be the case with attitude toward the church. In such situations, two people could give identical responses on the one-dimensional scale and yet be quite different as regards the two underlying processes. For instance, the same resultant could occur with two equal opposing tendencies of any given strength. Representing such data in the space formed by the underlying dimensions would yield a smear of points over an entire region rather than a simple curve as in Figure 2.

Although it may be reasonable to think that simple sensory phenomena are one-dimensional, it would seem that a considerable number of psychological variables must be conceived of as multidimensional in nature as, for instance, with "IQ" and other personality variables. Accordingly, as the two cited examples show, there is no logical guarantee that the use of ordinal scale statistics will yield invariant results under scale changes.

It is simple to construct analogous examples for nominal scales. However, their only relevance would be to show that a reduction of all results to categorical data does not avoid the difficulty with invariance.

It will be objected, of course, that the argument of the examples has violated the initial assumption that only "permissible" transformations would be used in changing the measuring scales. Thus, speed and time are not linearly related, but rather the one is a reciprocal transformation of the other. Similarly, Dimension 1 of Figure 2 is no monotone transformation of the original ordinal scale. This objection is correct, to be sure, but it simply shows that the problem of invariance of result with which one is actually faced in science has no particular connection with the invariance of "permissible" statistics. The examples which have been cited show that knowing the scale type, as determined by the commonly accepted criteria, does not imply that future scales measuring the same phenomena will be "permissible" transformations of the original scale. Hence the use of "permissible" statistics, although guaranteeing invariance of result over the class of "permissible" transformations, says little about invariance of result over the class of scale changes which must actually be considered by the investigator in his work.

This point is no doubt pretty obvious, and it should not be thought that those who have taken up the scale-type ideas are unaware of the problem. Stevens, at least, seems to appreciate the difficulty when, in the concluding section of his 1951 article, he distinguishes between psychological dimensions and indicants. The former may be considered as intervening variables whereas the latter are effects or correlates of these variables. However, it is evident that an indicant may be an interval scale in the customary sense and yet bear a complicated relation to the underlying psychological dimensions. In such cases, no procedure of descriptive or inferential statistics can guarantee invariance over the class of scale changes which may become necessary.

It should also be realized that only a partial list of practical problems of invariance has been considered. Effects on invariance of improvements in experimental technique would also have to be taken into account since such improvements would be expected to purify or change the dependent variable as well as decrease variability. There is, in addition, a problem of invariance over subject population. Most researches are based on some handy sample of subjects and leave more or less doubt about the generality of the results. Although this becomes in large part an extrastatistical problem (Wilk & Kempthorne, 1955), it is one which assumes added importance in view of Cronbach's (1957) emphasis on the interaction of experimental and subject variables. In the face of these assorted difficulties, it is not easy to see what utility the scale typology has for the practical problems of the investigator.

The preceding remarks have been intended to put into broader perspective that sort of invariance which is involved in the use of permissible statistics. They do not, however, solve the immediate problem of whether to use rank order tests or *F* in case only permissible transformations need be considered. Although invariance under permissible scale transformations may be of relatively minor importance, there is no point in taking unnecessary risks without the possibility of compensation.

On this basis, one would perhaps expect to find the greatest use of rank order tests in the initial stages of inquiry since it is then that measuring scales will be poorest. However, it is in these initial stages that the possibly relevant variables are not well-known so that the stronger experimental designs, and hence parametric procedures, are most needed. Thus, it may well be most efficient to use parametric tests, balancing any risk due to possible permissible scale changes against the greater power and versatility of such tests. In the later stages of investigation, we would be generally more sure of the scales and the use of rank order procedures would waste information which the scales by then embody.

At the same time, it should be realized that even with a relatively crude scale such as the rating scale of attitude toward the church, the possible permissible transformations which are relevant to the present discussion are somewhat restricted. Since the *F* ratio is invariant under change of zero and unit, it is no restriction to assume that any transformed scale also runs from 1 to 7. This imposes a considerable limitation on the permissible scale transformations which must be considered. In addition, whatever psychological worth the original rating scale possesses will limit still further the transformations which will occur in practice.

Although rank order tests do possess some logical advantage over parametric tests when only permissible transformations are considered, this advantage is, in the writer's opinion, very slight in practice and does not begin to balance the greater versatility of parametric procedures. The problem is, however, an empirical one and it would seem that some historical analysis is needed to provide an objective frame of reference. To quote an after-lunch remark of K. MacCorquodale, "Measurement theory should be descriptive, not proscriptive, nor prescriptive." Such an inquiry could not fail to be fascinating because of the light it would throw on the actual progress of measurement in psychology. One investigation of this sort would probably be more useful than all the speculation which has been written on the topic of measurement.

Question 3. Will the use of parametric as opposed to nonparametric statistics affect inferences about underlying psychological processes? In a narrow sense, Question 3 is irrelevant to this article since the inferences in question are substantive, relating to psychological meaning, rather than formal, relating to data reliability. Nevertheless, it is appropriate to discuss the matter briefly in order to make explicit some of the considerations involved because they are often confused with problems arising under the two previous questions. With no pretense of covering all aspects of this question, the following two examples will at least touch some of the problems.

The first example concerns the two students, P and Q, mentioned above, who had used time and speed as dependent variables. We suppose that their experiment was based on a 2×2 design and yielded means as plotted in Figure 1. This graph portrays main effects of both variables which are seen to be similar in nature in both panels. However, our principal concern is with the interaction which may be visualized as measuring the degree of nonparallelism of the two lines in either panel. Panel P shows an interaction. The reciprocals of these same data, plotted in Panel Q, show no interaction. It is thus evident in the example, and true in general, that interaction effects will depend strongly on the measuring scales used.

Assessing an interaction does not always cause trouble, of course. Had the lines in Panel P, say, crossed each other, it would not be likely that any change of scale would yield uncrossed lines. In many cases also, the scale used is sufficient for the purposes at hand and future scale changes need not be considered. Nevertheless, it is clear that a measure of caution will often be needed in making inferences from interaction to psychological process. If the investigator envisages the possibility of future changes in the scale, he should also realize that a present inference based on significant interaction may lose credibility in the light of the rescaled data.

It is certainly true that the interpretation of interactions has sometimes led to error. It may also be noted that the usual factorial design analysis is sometimes incongruent with the phenomena. In a 2×2 design it might happen, for example, that three of the four cell means are equal. The usual analysis is not optimally sensitive to this one real difference since it is distributed over three degrees of freedom. In such cases, there will often be other parametric tests involving specific comparisons (Snedecor, 1956) or multiple comparisons (Ducan, 1955) which are more appropriate. Occasionally also, an analysis of variance based on a multiplicative model (Williams, 1952) will be useful (Jones & Marcus, 1961). A judicious choice of test may be of great help in dissecting the results. However, the test only answers set questions concerning the reliability of the results; only the research worker can say which questions are appropriate and meaningful.

Inferences based on nonparametric tests of interaction would presumably be less sensitive to certain types of scale changes. However, caution would still be needed in the interpretation as has been seen in Question 2. The problem is largely academic, however, since few nonparametric tests of interaction exist.[3] It might be suggested that the question of interaction cannot arise when only the ordinal properties of the data are considered since the interaction involves a comparison of differences and such a comparison is illegitimate with ordinal data. To the extent that this suggestion is correct, a parametric test can be used to the same purposes equally well if not better; to the extent that it is not correct, nonparametric tests will waste information.

One final comment on the first example deserves emphasis. Since both time and speed are interval scales, it cannot be argued that the

[3] There is a nomenclatural difficulty here. Strictly speaking, nonparametric tests should be called more-or-less distribution free tests. For example, the Mood-Brown generalized median test (Mood, 1950) is distribution free, but is based on a parametric model of the same sort as in the analysis of variance. As noted in the introduction, the usual terminology is used in this article.

difficulty in interpretation arises because we had only ordinal scales.

The second example, suggested by J. Kaswan, is shown in Figure 3. The graph, which is hypothetical, plots amount of aggressiveness as a function of amount of stress. A glance at the graph leads immediately to the inference that some sort of threshold effect is present. Under increasing stress, the organism remains quiescent until the stress passes a certain threshold value, whereupon the organism leaps into full scale aggressive behavior.

Confidence in this interpretation is shaken when we stop to consider that the scales for stress and aggression may not be very good. Perhaps, when future work has given us improved scales, these same data would yield a quite different function such as a straight line.

One extreme position regarding the threshold effect would be to say that the scales give rank order information and no more. The threshold inference, or any inference based on characteristics of the curve shape other than the uniform upward trend, would then be completely disallowed. At the other extreme, there would be complete faith in the scales and all inferences based on curve shape, including the threshold effect, would be made without fear that they would be undermined by future changes in the scales. In practice, one would probably adopt a position between these two extremes, believing, with Mosteller (1958), that our scales generally have some degree of numerical information worked into them, and realizing that to consider only the rank order character of the data would be to ignore the information that gives the strongest hold on the behavior.

From this ill-defined middle ground, inferences such as the threshold effect would be entertained as guides to future work. Such inferences, however, are made at the judgment of the investigator. Statistical techniques may be helpful in evaluating the reliability of various features of the data, but only the investigator can endow them with psychological meaning.

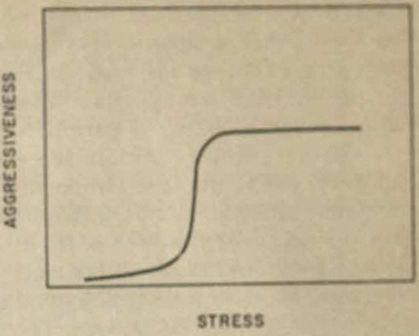


Fig. 3. Aggressiveness plotted as a function of stress. (The curve is hypothetical. Note the hypothetical threshold effect.)

Summary

This article has compared parametric and nonparametric statistics under two general headings: practical statistical problems, and measurement theoretical considerations. The scope of the article is restricted to situations in which the dependent variable is numerical, thus excluding strictly categorical data.

Regarding practical problems, it was noted that the difference between parametric and rank order tests was not great insofar as significance level and power were concerned. However, only the versatility of parametric statistics meets the everyday needs of psychological research. It was concluded that parametric procedures are the standard tools of psychological statistics although nonparametric procedures are useful minor techniques.

Under the heading of measurement

theoretical considerations, three questions were distinguished. The well-known fact that an interval scale is not prerequisite to making a statistical inference based on a parametric test was first pointed out. The second question took up the important problem of invariance. It was noted that the practical problems of invariance or generality of result far transcend measurement scale typology. In addition, the cited example of time and speed showed that interval scales of a given phenomenon are not unique. The discussion of the third question noted that the problem of psychological meaning is not basically a statistical matter. It was thus concluded that the type of measuring scale used had little relevance to the question of whether to use parametric or nonparametric tests.

REFERENCES

Anderson, N. H. Education for research in psychology. *Amer. Psychologist*, 1959, **14**, 695–696.

Boneau, C. A. The effects of violations of assumptions underlying the *t* test. *Psychol. Bull.*, 1960, **57**, 49–64.

Cochran, W. G. Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*, 1947, **3**, 22–38.

Cochran, W. G. Some methods for strengthening the common χ^2 tests. *Biometrics*, 1954, **10**, 417–451.

Coombs, C. H. A theory of psychological scaling. *Bull. Engrg. Res. Inst. U. Mich.*, 1952, No. 34.

Cotton, J. W. A re-examination of the repeated measurements problem. Paper read at American Statistical Association, Chicago, December 1959.

Cronbach, L. J. The two disciplines of scientific psychology. *Amer. Psychologist*, 1957, **11**, 671–684.

Dixon, W. J., & Massey, F. J., Jr. *Introduction to statistical analysis.* (2nd ed.) New York: McGraw-Hill, 1957.

Duncan, D. B. Multiple range and multiple *F* tests. *Biometrics*, 1955, **11**, 1–41.

Federer, W. T. *Experimental design.* New York: Macmillan, 1955.

Feldt, L. S. A comparison of the precision of three experimental designs employing a concomitant variable. *Psychometrika*, 1958, **23**, 335–354.

Gaito, J. Nonparametric methods in psychological research. *Psychol. Rep.*, 1959, **5**, 115–125.

Grant, D. A. Analysis-of-variance tests in the analysis and comparison of curves. *Psychol. Bull.*, 1956, **53**, 141–154.

Greenhouse, S. W., & Geisser, S. On methods in the analysis of profile data. *Psychometrika*, 1959, **24**, 95–112.

Jones, F. N., & Marcus, M. J. The subject effect in judgments of subjective magnitude. *J. exp. Psychol.*, 1961, **61**, 40–44.

Lindquist, E. F. *Design and analysis of experiments.* Boston: Houghton Mifflin, 1953.

Lord, F. M. On the statistical treatment of football numbers. *Amer. Psychologist*, 1953, **8**, 750–751.

Mood, A. M. *Introduction to the theory of statistics.* New York: McGraw-Hill, 1950.

Mosteller, F. The mystery of the missing corpus. *Psychometrika*, 1958, **23**, 279–290.

Rao, C. R. *Advanced statistical methods in biometric research.* New York: Wiley, 1952.

Savage, I. R. Nonparametric statistics. *J. Amer. Statist. Ass.*, 1957, **52**, 331–344.

Sawrey, W. L. A distinction between exact and approximate nonparametric methods. *Psychometrika*, 1958, **23**, 171–178.

Senders, V. L. *Measurement and statistics.* New York: Oxford, 1958.

Siegel, S. *Nonparametric statistics.* New York: McGraw-Hill, 1956.

Snedecor, G. W. *Statistical methods.* (5th ed.) Ames: Iowa State Coll. Press, 1956.

Stevens, S. S. Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology.* New York: Wiley, 1951.

Tate, M. W., & Clelland, R. C. *Nonparametric and shortcut statistics.* Danville, Ill.: Interstate, 1957.

Wilk, M. B., & Kempthorne, O. Fixed, mixed, and random models. *J. Amer. Statist. Ass.*, 1955, **50**, 1144–1167.

Williams, E. J. The interpretation of interactions in factorial experiments. *Biometrika*, 1952, **39**, 65–81.

(Received April 8, 1960)

Psychological Bulletin
1961, Vol. 58, No. 4, 317–324

BASIC FORMS OF COVARIATION AND CONCOMITANCE DESIGNS

RICHARD W. COAN
University of Arizona

Several years ago, Cattell (1946) published a description of what he called the "covariation chart," a graphic model which illustrates six basic forms of covariation with which we may deal in psychological research. It is the purpose of the present paper to describe an extension and modification of Cattell's schema that will provide much more comprehensive classification of actual and possible research designs in psychology.

The six forms of covariation encompassed by Cattell's model have been variously labeled with letters through the alphabetic range from M to T, but the labeling indicated in Table 1 has come to be reasonably standard. The covariation chart itself consists of a parallelepiped, in which the three dimensions represent tests, persons, and occasions. Any plane parallel to any surface of the model represents a score matrix which might correspond to the data from a psychological research. There are three such sets of planes, any one plane permitting consideration of two kinds of covariation.

TABLE 1

THE SIX BASIC FORMS OF COVARIATION INDICATED IN THE COVARIATION CHART

Technique	Variables correlated	Series over which correlated	Variables held constant or singular
R	Tests	Persons	Occasions
Q	Persons	Tests	Occasions
P	Tests	Occasions	Persons
O	Occasions	Tests	Persons
S	Persons	Occasions	Tests
T	Occasions	Persons	Tests

The major virtue of a classification scheme like that embodied in the covariation chart is that it can suggest forms of valuable research which might otherwise be overlooked. As Cattell himself has clearly recognized, however, the scope of the covariation chart model has certain unfortunate limitations. When he first presented the covariation chart, Cattell pointed out that the six techniques did not really exhaust the forms of covariation inherently derivable from the three-dimensional model. The other forms which he considered at that time, however, are essentially variants or compounds of the six basic forms of covariation.

Various more novel techniques will emerge, of course, if we can find justification for adding other dimensions to the model. In a more recent publication, Cattell (1957) points out that a psychological event may be characterized in terms of six independent "tags": a reacting organism, a focal stimulus, a background condition, a response, an occasion in time and space, and an observer. He suggests that any pair of tags may serve as the dimensions of a score matrix yielding a technique and its transpose. Since there are 15 possible pairs of tags, there are 15 possible techniques (and their corresponding transposes). Furthermore, the elements within any matrix could correspond to any of the six tags. Logically, this would extend the system to 90 possible techniques (or 180, including transposed techniques). Cattell apparently excludes some combinations and speaks of 45 possible techniques. To these he adds five additional possibilities that involve a mixture of tags along one axis of the score matrix.

In the view of the writer, the original covariation chart provides too limited a classification system. The extended model, however, introduces needless complexity and is subject to useful modification and simplification. Cattell's six tags represent six distinguishable aspects of any observed psychological event, but they do not, on that account, constitute six meaningfully distinguishable aspects of research design.

The distinction between focal stimulus and background condition is a somewhat arbitrary one, and its usefulness in design classification is questionable. We can nearly always isolate a great variety of stimulus variables that will influence a given event in a more or less direct way. Insofar as the researcher analyzes the effect of one of these variables, it becomes a focal stimulus variable, at least from the standpoint of the researcher and hence of the research design. Background conditions are otherwise irrelevant to experimental design, unless they are confounded with other kinds of variables (organisms or occasions).

The observer is also a vital part of any psychological event dealt with in research, but the observer becomes important as a component of design only to the extent that he is something more than an observer. If his presence in the situation affects the behavior of the subject of the experiment, the observer becomes to that extent a part of the stimulus situation and may be analyzed accordingly. If our interest, on the other hand, is in peculiarities of the observer as a recorder or rater of behavior, we are to that extent treating him as a reacting organism, i.e., as the subject of an experiment superimposed on another experiment.

Basic Components of Design in Psychological Research

It is possible to characterize a psychological event in terms of a great number of distinguishable features which set it apart from other psychological events, but there are basically only four such features that constitute essential and distinguishable parameters of any research design employed to study psychological events. We shall refer to these features henceforth as *design components* and label them *R*, *S*, *P*, and *O* (not to be confused with *Techniques R*, *S*, *P*, and *O*).

Design Component *R* is that realm of variables which consists of structural or functional manifestations on the part of the subject or subjects under investigation and which are studied through observation and measurement of the subject or of products of the subject's behavior. Commonly treated as single *R*-component variables are specific responses, score summaries of patterns or sets of responses, and attributes. Design Component *S* is that realm of variables which arises from sources outside the subject and which may be expected to influence the subject's behavior. *S*, then, refers to external stimuli. Those things which are sometimes called "internal stimuli" fall within the scope of *R*-component variables if they are directly observed or measured. The *P* component is that of the human or animal subjects observed in the experiment. The *O* component is the realm of occasions, in given time and space, on which experimental observations are made.

These four components are ordinarily quite distinct from one another and subject to separate specification. For some purposes, we may artificially tie variables of one component to those of another. In such cases, we may speak of a "confounding" of design components. Confounding is most common with respect to Component *O*, which for various purposes we permit to vary systematically with certain *S*, *P*, or *R* variables. A confounding of *S* and *P* variables is also quite common.

In a sense, any variable that we observe and describe may be said to be measured, at least implicitly, for if our description contains only an identifying qualitative statement, we have provided the essential ingredients of nominal scaling. Since the variables of all four design components are subject to observation and description in a psychological experiment, they may be regarded as subjected simultaneously and independently to measurement and scaling. Within any component, variables may be scaled at any level—nominal, ordinal, interval, or ratio—and are sometimes simultaneously scaled at more than one level.

A peculiarity of Component *P* that should be noted is that data within it are usually treated as scaled either at the nominal or at the ratio level. So long as we are concerned merely with identifying individuals as distinguishable entities, we make only the assumptions of nominal scaling. When we treat individuals as equivalent units that can be added together, however, and express *P*-component data in terms of numbers of cases or proportions of a total sample of subjects, we have made the essential assumptions underlying ratio scaling. The data could be expressed in ordinal form if the label identifying the individual assumed the form of an index of rank within a social hierarchy. We could transform the data from ordinal form to presumably interval form either by making certain parametric assumptions or by adopting some appropriate measure of discriminability of adjacent ranks as an index of interval size. (Numerical data within the realm of Component *P* may assume any form consistent with the notion of *measurement in terms of individuals.* The application of measurement *to* individuals, however, yields *R*-component data.)

An Extended Covariation Design Classification

A consideration of the role played by variables of the four design components in the covariation chart reveals that *R*-component variables are consistently assigned to the cells within the score matrices corresponding to Techniques *R*, *Q*, *P*, *O*, *S*, and *T*. The numbers in the body of a score matrix represent what we conceive of as the dependent variable in an experiment. In psychological research, the dependent variable is customarily, but not inevitably, the response variable. While our interest may lie in finding what sort of response will appear in a given situation, we may seek, with equal justification, to determine which individual will give a particular response, which stimulus will evoke the response, or on what occasion the response will appear. If we thus permit any of the four design components to furnish the elements within the score matrix, we are led to the system of 24 techniques shown in Table 2.

It may be noted that no component appears twice in any row of Table 2.

TABLE 2

An Extended System of Covariation Designs

Technique	Variables correlated	Variable in which variation is noted	Series over which covariation is studied	Constant or singular variable
R	*S*	*R*	*P*	*O*
Q	*P*	*R*	*S*	*O*
P	*S*	*R*	*O*	*P*
O	*O*	*R*	*S*	*P*
S	*P*	*R*	*O*	*S*
T	*O*	*R*	*P*	*S*
A	*R*	*S*	*P*	*O*
B	*P*	*S*	*R*	*O*
C	*R*	*S*	*O*	*P*
D	*O*	*S*	*R*	*P*
E	*P*	*S*	*O*	*R*
F	*O*	*S*	*P*	*R*
G	*R*	*P*	*S*	*O*
H	*S*	*P*	*R*	*O*
I	*R*	*P*	*O*	*S*
J	*O*	*P*	*R*	*S*
K	*S*	*P*	*O*	*R*
L	*O*	*P*	*S*	*R*
U	*R*	*O*	*S*	*P*
V	*S*	*O*	*R*	*P*
W	*R*	*O*	*P*	*S*
X	*P*	*O*	*R*	*S*
Y	*S*	*O*	*P*	*R*
Z	*P*	*O*	*S*	*R*

Note.—The letters in the second, third, fourth, and fifth columns refer to the design components from which variables are drawn.

This classification system assumes that the two axes of the matrix and the elements within the matrix will generally represent three different design components. Supporting this assumption is the fact that each design component represents variables which are an integral part of any psychological event, and the questions raised in psychological research normally refer to the manner in which variables of the different realms represented by the four components converge in a given psychological event. It must be granted, however, that our assumption is, in some respects, an arbitrary one. It is possible to conceive of designs in which the axes and the matrix elements would not represent three different components, but such designs can also be rationalized quite readily as variants of techniques already in thesystem. Whether the classification system proposed here will generally provide the most convenient framework for design conceptualization must ultimately be determined through practical application. In any case, a classification system of this sort cannot be exhaustive if it is to remain fairly simple. It can merely provide a framework of basic prototypal techniques. Some designs will inevitably appear as combinations or variants of these techniques.

It must be emphasized that these techniques refer to research designs in which covariation is to be observed, but they do not imply any particular form of statistical analysis. In general, the desired indices of covariation will be furnished by correlational methods. Whether a method such as factor analysis or cluster analysis will be applied subsequently is an additional consideration.

Covariation Design and Concomitance Design

If we are interested in truly comprehensive classification of psychological research designs, we must recognize at the outset that most psychological experiments are not actually concerned with covariation. The simplest form of research would call for a single measurement. This measurement might fall within the realm of any of our four design components, and it could be thought of as the single element filling a single-cell matrix. The variables of the other three components would also be singular.

More commonly we speak of research design when we seek data for a matrix of at least two cells and where we are interested in a relationship among the ingredients of the matrix. The relationship may nearly always be considered in terms of a concomitance of two or more elements falling within the realm of one of our design components, and these elements are related in terms of their convergence with elements corresponding to a different component. If we represent all variables or elements of a common design component along a common axis of a score matrix, the data of many experiments must be thought of as filling cells arranged serially in a single row or column. We relate either the single cell rows of a single column matrix or the single cell columns of a single row matrix.

The kind of matrix we are now describing is a truncated version of the kind we assumed in classifying covariation designs. We can speak meaningfully of *concomitance* with respect to two single cell rows, but not of *covariation*, for this assumes two relatable series of values. In a single column matrix, whatever component would otherwise have constituted a horizontal axis is now treated as singular.

Nearly every psychological research design is concerned with concomitance, but not necessarily with covariation (i.e., concomitant variation). Since the covariation design is really a special case of concomitance design, it would be worthwhile to have a scheme of classification for concomitance designs which would parallel that for covariation designs. Such a scheme is presented in Table 3. Since in each concomitance design the serial variable is replaced by an additional singular variable, each concomitance design may be considered a truncated version of either of two covariation designs.

TABLE 3

BASIC CONCOMITANCE DESIGNS

Technique	Variables related	Variables in which variation is noted	Singular or constant variables	Parallel covariation designs
Alpha	*S*	*R*	*P*, *O*	*R*, *P*
Beta	*P*	*R*	*S*, *O*	*Q*, *S*
Gamma	*O*	*R*	*S*, *P*	*O*, *T*
Delta	*R*	*S*	*P*, *O*	*A*, *C*
Epsilon	*P*	*S*	*R*, *O*	*B*, *E*
Zeta	*O*	*S*	*R*, *P*	*D*, *F*
Eta	*R*	*P*	*S*, *O*	*G*, *I*
Theta	*S*	*P*	*R*, *O*	*H*, *K*
Iota	*O*	*P*	*R*, *S*	*J*, *L*
Kappa	*R*	*O*	*S*, *P*	*U*, *W*
Lambda	*S*	*O*	*R*, *P*	*V*, *Y*
Mu	*P*	*O*	*R*, *S*	*X*, *Z*

Note.—The letters in the second, third, and fourth columns refer to design components from which variables are drawn.

APPLICATIONS OF CONCOMITANCE DESIGNS

The techniques labeled Alpha, Beta, and Gamma in Table 3 represent the most familiar forms of psychological research, and in them we find the most frequent application of such forms of statistical analysis as the critical ratio and analysis of variance. Beta technique has a common application in the comparison of responses of groups which differ with respect to variables outside the range of observation within the experiment (e.g., two different occupational groups, psychotics and "normals," men and women, etc.). Comparison of matched groups subjected to different stimulus conditions would constitute a form of Alpha technique, since *P*-component variables are held constant. Interest is here focused on the relating of stimuli, as in the simpler form of Alpha technique involving such a comparison for a single individual or single group of individuals. A *compounding* of techniques is possible in designs of more than one-way classification. Thus, we should have a compound of Alpha and Beta techniques if we classified both in terms of known group membership and in terms of stimulus conditions. The reader will note that the score matrix in terms of which we conceptualize the design differs from the tabular arrangement usually employed with analysis of variance in that the variables to be related are represented along a common axis. Thus the score matrix for a complex factorial design of the Alpha-technique variety would consist of a long single column of *R* data. Each row would represent the data for a group simultaneously scaled with respect to several stimulus dimensions.

In Techniques Delta, Epsilon, and Zeta, the stimulus is conceptually the dependent variable. These techniques bring to mind certain applications of psychophysical methods. Strictly speaking, the procedures usually called "psychophysical methods," as described by such writers as Graham (1950) and Guilford (1954), are methods of measurement and do not define specific experimental designs to any greater extent than do methods of statistical analysis. In actual application, however, they form a basis for a limited range of concomitance designs.

The most common applications of psychophysical methods may be thought of as constituting either Alpha technique or Delta technique, depending largely on the use made of the data. The simple application of the method of constant stimuli, for example, would constitute Delta technique if we dealt with the resulting data in terms of a relationship between the two response categories. Each of the two cells of the corresponding score matrix would contain the value of the stimulus eliciting the given response for a certain percentage of trials. On the other hand, findings may be expressed by means of a curve in which stimulus magnitude is plotted against the percentage of trials in which either response is produced. The design may then be considered either Alpha technique or Delta technique, depending on whether we consider the curve as a way of expressing relationships within a continuous series of *S* categories or within a continuous series of *R* categories (percentages in the present instance). Similar reasoning would apply, of course, to the application of other psychophysical methods. More complex applications of these methods, in which *R* variables are related to a combination of interacting *S* dimensions—as in Licklider's (1951) treatment of auditory functions—may be regarded as comparable to the application of factorial design in Alpha technique. Psychophysical methods are less commonly applied in research classifiable as Epsilon or Zeta technique, although certain applications of these methods in clinical research (e.g., certain studies involving flicker fusion, size judgments, distance judgments, and judgments of the vertical) would certainly qualify as Epsilon technique.

Techniques Eta, Theta, and Iota are a common realm of application for nonparametric techniques of statistical analysis. Depending on the manner in which *P*-component data are expressed, we may analyze findings in terms of cell frequencies, overlap of cases among cells, or comparability of person ranks associated with various cells.

Techniques Kappa, Lambda, and Mu are most likely to be useful when variations in an occasion variable are presumed to covary with certain attributes of subjects or with certain changes in the life situations of subjects. The *O*-component variable may thus reflect such things as age, developmental stage, and level or stage of experience. The developmental area is probably the most common realm of application. Kappa technique provides a means for grouping behaviors developmentally and hence for defining developmental stages. Lambda technique provides a way of defining stages in terms of effective stimuli. Mu technique can be used to compare individuals with respect to such things as rate of maturation. Many applications outside the developmental realm, to processes involving shorter time spans, are possible.

Applications of Covariation Designs

A detailed discussion of possible applications of the familiar Techniques *R*, *Q*, *P*, *O*, *S*, and *T* would be superfluous here. Unfortunately, other treatments of these techniques have promoted misconceptions by obscuring three interrelated considerations that are basic to consistent classification. First, there is the distinction between concomitance and covariation designs. A second vital point is that the series over which covariation is observed must be genuinely treated as a series in covariation designs. Wherever a group is treated as a unit and a group average is treated as a single observation, the group functions, for design classification purposes, as a single individual. Finally, in research employing matched groups, the *P* component is properly viewed as being held constant, and appropriate classification will depend on what component is confounded with Component *P*. For example, in the common type of experiment in which equated control and experimental groups are subjected to different stimulus conditions, we have an instance of *P* technique (not *S* technique, as some writers would have it), provided that response covariation is considered over time. The usual application of this design, to a single occasion, is simply Alpha technique.

The remaining techniques—*A* through *L* and *U* through *Z*—represent virtually unexploited forms of design, but careful consideration will suggest appropriate uses for each of them. In Techniques *A* through *F*, the dependent variable is of Component *S*. Appropriate quantification might be in terms of minimally sufficient stimulus magnitude or mean stimulus magnitude associated with a given response. In Techniques *G* through *L*, the dependent variable is of Component *P*. It may be expressed in terms of the rank of the individual giving a certain response to a certain stimulus on a certain occasion, in terms of the average rank of individuals so responding, or in terms of the number of individuals so responding. In Techniques *U*, *V*, *W*, *X*, *Y*, and *Z*, our focal variable—of Component *O*—may be expressed in terms of a single occasion in an ordered series, an average of a number of ordered occasions, an average age, an average stage, etc. It is important to note that in covariation designs, in contrast to concomitance designs, the dependent variable must be of at least the ordinal level of scaling. Thus, in Techniques Eta, Theta, and Iota, the matrix cells could simply contain tags identifying the persons fitting the cell coordinates. Data analysis would then consist of assessing the overlap of entries in various cells. In a matrix of several rows and columns containing such nominal data, we could probably speak of "multiple concomitance" with respect to a pair of rows or columns, but it is doubtful that we can properly speak of "covariation" unless the data in the body of the matrix are expressed in a form representing relative magnitudes or positions on continua.

Techniques *A*, *C*, *G*, *I*, *U*, and *W* all deal with the covariation of response categories and thus provide a basis for defining the structure of a response realm. In *C* and *U*, we assess the comparability of response categories on an intra-individual basis, with respect to conjoint appearance on various occasions or in response to various stimuli. Techniques *A*, *G*, *I*, and *W* provide a means for assessing response comparability on a group basis, in terms of similarity of the precipitating stimulus, of the occasion of manifestation, or of the persons giving the response.

Techniques *R* and *P* are familiar techniques for examining stimulus covariation in terms of the resulting response. Techniques *H*, *K*, *V*, and *Y* add possibilities for correlating stimuli in terms of covariation with respect to the magnitudes (ranks) or numbers of persons responding in a certain way or in terms of the particular occasions or numbers of occasions on which the stimulus has a given effect. The correlating of persons is also a familiar idea by virtue of its introduction through *Q* and *S* techniques. Techniques *B*, *E*, *X*, and *Z* point to the possibility of correlating persons according to stimuli producing various responses, stimuli producing a given response on various occasions, occasions when various responses appear, or occasions when various stimuli elicit a given response. Consideration of the many possible ways of defining the basic stimulus, response, and occasion data suggests a great variety of ways of grouping persons according to such things as physiological cycles, social roles, and developmental patterns.

In applying any of the occasion-correlation techniques—*O*, *T*, *D*, *F*, *J*, and *L*—we may select presumably equivalent occasions and thus obtain an estimate of the reliability, or stability, of a given pattern of relationship. We may, on the other hand, select occasions differing in a known way and thereby determine the comparability of these occasions. Possible applications range from the psychophysical realm to the developmental realm, depending on how the occasion variable is defined and quantified. In general, the new covariation techniques encompassed by this expanded classification system promise a rich harvest through novel approaches to diverse problems—particularly in the developmental, social, and physiological areas, where the possible fruits of correlational analysis have been recognized by too few researchers.

REFERENCES

CATTELL, R. B. *The description and measurement of personality.* New York: World Book, 1946.

CATTELL, R. B. *Personality and motivation structure and measurement.* New York: World Book, 1957.

GRAHAM, C. H. Behavior, perception, and the psychophysical methods. *Psychol. Rev.*, 1950, **57**, 108–120.

GUILFORD, J. P. *Psychometric methods.* New York: McGraw-Hill, 1954.

LICKLIDER, J. C. R. Basic correlates of the auditory stimulus. In S. S. Stevens (Ed.), *Handbook of experimental psychology.* New York: Wiley, 1951. Pp. 985–1039.

(Received April 28, 1960)

Psychological Bulletin
1961, Vol. 58, No. 4, 325–336

THE SELF-CONCEPT: FACT OR ARTIFACT?

C. MARSHALL LOWE[1]
Ohio State University

One of the more difficult tasks for psychology is relating the observation of behavior to the study of mental processes. One approach to the problem has been to limit psychology to the study of behavior and to leave to philosophy the task of speculating as to the existence and nature of mind and soul.

There have, however, been psychologists who have sought to make sense out of human action by positing a self or ego, in order that they might understand the coherence and unity which they have thought that they have seen in human behavior. Thus, G. W. Allport (1943) claimed that the concept of ego was made necessary by certain shortcomings in associationism, and he went on to list eight different uses for the concept of the ego. During the 1940s the *Psychological Review* was in fact well-flavored with articles of philosophical taste (Allport, 1943; Bertocci, 1945; Chein, 1944; Lundholm, 1940). These articles were attempts to find the source of human behavior by discussions of concepts, but they failed to make a lasting distinction between the self as subjective knower and the self as object of knowledge. The self as essence defied definition, and the discussions concerning the nature of mind seemed relevant for neither experimental nor applied psychology.

But during the 1940s there was a parallel attempt at construction of a useful concept of the self. While Rogers wrestled with the problem of researching a client centered approach in psychotherapy, one of his students (Raimy, 1943) developed a construct of the self which had a perceptual frame of reference. What Raimy called the self-concept was both a learned perceptual system functioning as an object in the perceptual field, and a complex organizing principle which schematizes ongoing experience. Raimy demonstrated in his dissertation that attitudes toward the self can be found by analyzing counseling protocols, and that these self-perceiving attitudes formed a reliable index for improvement in psychotherapy.

The concept of the self soon formed the theoretical underpinning for a new approach to the study of behavior. Raimy's construct of the self received further development in the book *Individual Behavior* (Snygg & Combs, 1949). The authors stated that behavior was best understood as growing out of the individual subject's frame of reference. Behavior was to be interpreted according to the phenomenal field of the subject rather than be seen in terms of the analytical categories of the observer.

As the self-concept was born with client centered therapy, so congruent were the theory of the self and the practice of psychotherapy that a new self centered therapy became theoretical for the first time: Rogers (1951) described therapeutic change in a phenomenological frame of reference.

[1] Now at Maryville College.

The author is indebted to Collins W. Burnett of the Department of Psychology for his encouragement and suggestions in the preparation of this manuscript.

By 1950 the phenomenological view of the self had become the center of a new movement in psychology, having already generated a block of research studies (Rogers et al., 1949). When Hilgard (1949) postulated in his APA presidential address the need for a self to understand psychoanalytic defense mechanisms, and called for research on the self, psychology listened. To the desert came rain that washed all before it.

The deluge of studies within the last decade has not been contained within any one theoretical channel, so that studies involving the self-concept have spread over into many areas of psychology. Ten years of research efforts have produced a mass of data, reflecting different theoretical assumptions and differing research methods. While the time has now passed for one article to deal adequately with all the studies that have been done, the sheer mass of evidence would suggest that certain questions be asked of theories of the self-concept.

This paper is concerned with the problem as to whether the self is an objective reality which is a fit field for psychological research, or whether it is a somewhat nebulous abstraction useful only to give a theoretical basis to things the psychologist could not otherwise understand. Put in other words, this paper faces the issue as to whether the results of studies of the self are to be accepted at face value, or whether other explanations of results would be more parsimonious or reasonable.

The writer will discuss first attempts to quantify data concerning the self-concept to arrive at an operational definition. We will then assess the validity of measures of the self-concept, and will relate the self-concept to other constructs. We will briefly allude to attempts to establish a relationship between different measures. Finally, the writer will return to certain philosophical and historical considerations in order to reach a conclusion as to whether the self-concept is indeed a fact of nature, or an artifact of men's minds.

Measuring the Self-Concept

Many psychologists have believed that if something exists it can be measured. There have been many investigators who have assumed that the self-concept refers to an existence of some sort and have gone on to measure it.

The most popular type of operational definition has assumed that the self-concept can be defined in terms of the attitudes toward the self, as determined either by the subject's references to himself in psychotherapy or by asking him to mark off certain self-regarding attitudes on a rating scale.

One of the first attempts at attitude measurement was by Sheerer (1949), who extracted from the protocols of cases at the University of Chicago Counseling Center all statements that were relevant either for attitudes to self or to other people. These statements formed the basis for a 101-item rating scale. The Sheerer client statements also formed the basis for rating scales constructed by Phillips (1951) and by Berger (1952).

The only rating scale of attitudes towards self that has been published is the Index of Adjustment and Values (Bills, 1958). Bills states that the intent of the index is to measure the phenomenological self view as described by Lecky (1945), Snygg and Combs (1949), and Rogers (1951). This scale is more elaborate in that each item is ranked with three different instructions. First, the subject ranks the item on a scale as to how well it describes himself. Next, he marks the items as to how acceptant he is of his first, or self-rating of the item, and finally he rates the item as to the degree to which he aspires to be like that item.

The scoring of the Bills index also is more elaborate than that traditional for rating scales. There are in fact two different measures, neither one being simply a rating of items in absolute terms, as in the scales previously described. Bills' measures depend instead upon the differences between ratings made under different instructions. A measure of self-acceptance is provided by the degree of similarity between the way the subject sees himself as being, and the way he rates himself as accepting his self-ratings. A measure of self–ideal-self discrepancy is given by comparing the differences in ratings between the way the self is rated as being, and the way the self is rated as wishing to be.

Brownfain (1952) made still another adaptation in the use of the rating scale, deriving a measure of what he termed the stability of the self-concept. Subjects ranked themselves on 25 words and phrases, each describing a different area of personality adjustment. The measure is not of how sure the subject is of himself, but of how sure he is of what he thinks about himself; the subject is instructed to make the ratings twice, first with an optimistic frame of reference, and then with a pessimistic one. The degree of congruence between the two ratings is termed the degree of stability of the self-concept.

A different theoretical approach towards measurement of self-concept involves the use of *Q* technique. Stephenson (1953) describes how one's "inner experiences" can be translated into behavior by means of *Q* sort, through which the phenomenal field is translated into action. Using this method, two of Stephenson's students at the University of Chicago derived a conceptual self-system in an intensive study of a single subject (Edelson & Jones, 1954).

Others at the University of Chicago have used *Q* sorts as a measure of self-concept, in an attempt to assess changes in self-concept during psychotherapy (Rogers & Dymond, 1954). Statements were taken from counseling protocols, and were sorted both for real self and for ideal self. The degree of congruence between the two sorts is taken as a measure of adjustment.

Attempts to measure the self-concept face three difficulties. First, it must be demonstrated that the operational and philosophic meanings are in fact equivalent. In the case of the self-concept it needs to be shown that the "inner experience" is effectively conveyed by the outward movement of making check marks on lines, or sorting cards. Secondly, an efficient and systematic method must be found for selecting items for the scales and sorts, the problem being that of defining the universe from which items are to be selected. Finally, the different measures imply different operational definitions. Just as one can not multiply apples and pears, so is it impossible to interchange different operational definitions as if they were the same, or to pretend that each means the same thing by the term self-concept.

If something is measured does it exist? If the answer is yes, we must still be aware that we may not fully understand what we are measuring.

One must measure, but must then compare and carefully validate.

Validation of Self-Concept Measures

A psychological construct stands and falls according to how useful it is in understanding human behavior. A term is meaningful only when successful validation studies have found significant relationships with established variables.

It has been popular to validate self-concept scales against tests purporting to measure maladjustment in an attempt to demonstrate that one's phenomenological view of the self is closely related to the degree of adjustment. Positive results abound. Calvin and Holtzman (1953) had college students rank themselves on seven personality traits, and found that self-depreciation was related to high scores on the MMPI. Zuckerman and Manashkin (1957) had neuropsychiatric patients rate themselves on a scale of adjectives, and found that self-ratings correlated positively with the MMPI *K* scale, and negatively with seven of the other scales. Taylor and Combs (1952) tested the hypothesis that sixth grade children found to be well-adjusted on the California personality scale would more often admit statements of self-reference which though unflattering were universally true. They got positive results, the self-depreciation which in other self-concept measures is treated as vice being here treated as virtue. Hanlon, Hofstaetter, and O'Connor (1954) compared the results of high school juniors on the California personality scale with the degree of congruence between ratings of the real and ideal self and found that the more congruence the better the adjustment. Cowen (1954) related low self-ratings on the Brownfain negative self-concept with high scores on the California F Scale. Any doubt about the ability of investigators to find positive results when comparing good adjustment as measured by objective personality inventories with the affirmativeness of self-concept should be dispelled by a study by Smith (1958). He compared congruence between *Q* sorts for self and ideal self with scores on the Edwards PPS, the Cattell factors, and measures of average mood. After making almost 300 correlations, he concluded that having a positive self-concept is indeed related to adjustment.

Other investigators have doubted that the relationship between adjustment and self-satisfaction is such a simple one. Block and Thomas (1955) conceived of maladjustment lying at both ends of the continuum. They felt that too high a degree of self-satisfaction is due to suppressive and repressive mechanisms which cause a person to be rigid, over-controlled, restrained, and aloof. But at the other extreme, the person who is too little satisfied with self will lack ego defenses, and will be able neither to bind tensions nor control emotions. Block and Thomas constructed an ego-control scale from MMPI items. The scale was found to have a correlation of .44 with self–ideal-self *Q* sort congruence, the relationship being curvilinear. Unfortunately, this was the reverse of what Chodorkoff (1954a) had found. Correlating ratings of the self as made from a biographical inventory with the results of projective techniques, he found that maladjustment lies in the middle range of self-satisfaction.

Validating self-concept measures against objective personality tests has generally been successful, but the true significance of these studies is still not made clear. Edwards (1957) demonstrates how more than half the variance in both MMPI scales and in *Q* sorts of self-referent items is accounted for by social desirability. SD can account for significant positive relationships even when other variables are totally unrelated. Edwards' SD robs these studies neither of significance nor of interest, but does suggest that extreme care must be taken in the labeling of constructs.

Attempts have also been made to validate self-concept against projective personality tests. Bills has made several attempts to validate his scale by the Rorschach (Bills, 1953a, 1954; Bills, Vance, & McLean, 1951). The results are a bit ambiguous, and leave two observers (Cowen & Tongas, 1959) extremely dissatisfied. The TAT was used by Friedman (1955) to compare the *Q* sort discrepancy self with the self as projected onto the TAT pictures. The normals were the only group to project positive self-qualities. Neurotics and paranoids both projected negatively.

A different approach to validation has used a word association test. Results show that there is a delayed reaction time for those trait words where there has been a discrepancy in ratings between the self and the ideal self (Bills, 1953b, Roberts, 1952). Delayed associations are assumed to be related to defensiveness about self, which in turn is considered to be related to maladjustment. However, Cowen and Tongas (1959) wonder if defensiveness about trait words does not serve also to raise the original ratings of the actual self.

Cowen chose to validate the self-concept by comparing the absolute self-rating with the learning time for the rated words, and found that there was a higher learning time for words that were presumably threatening. We might however wonder if his and Tongas' criticism of other studies does not apply here also: defensiveness might also cause self-ratings to be raised.

Use was also made of the perceptual New Look. Chodorkoff (1954b) presented neutral and threatening words with a tachistoscope, and found that the better the agreement between a self-description and a description of the self by others, the less perceptual threat there will be.

It is unfortunate that the only study of this general type that did not use college students as subjects was negative in its results. Zimmer (1954) presented male mental patients with trait adjectives on which there was a self-rating discrepancy between self and ideal self. A word association test was not found to be significantly related to self-discrepancy.

The results of studies that involve the presentation of "hot" or threatening words seem suggestive, for there seems to be a common element in ability to free associate and learn threatening words. But it is possible that we have in these studies more a measure of ego defenses than of maladjustment, for the fact that the results are positive only with normal groups might suggest that the results are more relevant for a theory of personality than for a theory of psychopathology. In these studies it is indeed likely that we have support for Lecky's theory of self-consistency, and for Snygg and Comb's theory of the maintenance of internal organization. If this is so, then likely it is true as Block and Thomas (1955) suggest that only extremes in ego control are pathological.

A different approach to validation

of self-concept measures uses behavior in a social situation as a criterion. The most sweeping results in a study of this type are reported by Turner and Vanderlippe (1958), who report that *Q* sort congruence between the self and the ideal self is greater in those college students who are more active in extracurricular activities, have higher scholastic averages, and are given higher sociometric rankings by fellow students. Holt (1951) found that agreement between self-ratings and ratings by a diagnostic council was positively related to intelligent, active, adventurous living, and a friendly dominant social adjustment. Eastman (1958) found that the degree of acceptance of self-ratings on the Bills index is positively related to ratings for marital happiness. Working in terms of ratings for maladjustment, Chase (1957) found that among maladjusted patients there was greater discrepancy between *Q* sorts for self as compared with sorts for the ideal self and the average other person.

Other attempts to relate self-concept to social behavior have been less successful. Kelman and Parloff (1957) obtained only chance results when they tried to interrelate such variables as congruence between self and ideal self, a symptom disability check list, a discomfort evaluation scale, sociometric ratings, and an ineffective behavior evaluation scale, using 15 neurotic hospital outpatients. Fiedler, Dodge, Jones, and Hutchins (1958) measured the self-concept of college students both by a simple rating scale and by a discrepancy measure. There was a general lack of correlation between these measures and such objective criteria as grade point average, health center visits, army adjustment, the Taylor *MA* scale, and sociometric status. Coopersmith (1959) compared self-esteem as rated by the self with that estimated by observers, using children as subjects. He suggests that there are actually four types of self-esteem: what a person purports to have, what he really has, what he displays, and what others believe he has.

There is no obvious explanation for the discrepancy of results in studies purporting to relate self-concept to behavioral adjustment. Since the basis for selecting items for rating scales and *Q* sorts differs from study to study, it is possible that the statements used in the scales of those studies with positive results had more of a relationship to the criteria than the statements in studies which were negative.

A different approach to relating self-concept measures to adjustment is shown in a block of psychotherapy research studies at the University of Chicago (Rogers & Dymond, 1954). Change in self-concept was found to occur as a function of improvement during psychotherapy. Butler and Haigh (1954) had clients make *Q* sorts for self and for ideal self both before therapy and after its completion to test the hypothesis that therapy will increase satisfaction with the self. Congruence between the two sorts increased as a result of psychotherapy, the two sorts moving towards a common mean. Rudikoff (1954), using the same subjects, found changes during periods of time before and after therapy were not nearly as great as those occurring during therapy. Also with the same subjects, Dymond (1954) found that there was closer agreement after therapy between the way clients sorted the Butler and Haigh *Q* sort cards, and the way two non-Rogerian clinical psychologists sorted the cards between what the well-adjusted person should say is like him and what is not like him.

The same investigator (Cartwright, 1958) related change in self-concept over therapy to a successful search for identity. She had clients make sortings with Butler and Haigh *Q* sort cards to describe themselves as they saw themselves in relationship to three people of their choice to test the hypothesis that successful therapy increases the consistency of the self-concept which one brings to different social situations. The hypothesis was confirmed.

Ewing (1954) had counselee college students rate a list of traits for self, ideal self, mother, father, counselor, and a culturally approved figure. There was a regression of the ratings toward a common mean in those clients who were estimated to be the most improved in therapy.

Changes in self-ratings over therapy seem certainly to have occurred. But they seem to take place also without psychotherapy. Taylor (1955) devised a *Q* sort divided between positive and negative statements. After subjects made repeated sortings both for self and for ideal self, he concluded that self-introspection without therapy results in increased positiveness of attitude toward the self; that the self and ideal self will draw closer together; and that repeated self-descriptions are accompanied by increased self-consistency. Engel (1959) studied the stability of self-concept in adolescence, and also found a trend towards more positive *Q* sorting over a 2-year period. And finally Dymond herself (1955) found an increased congruence between *Q* sorts for self and for ideal self among subjects waiting for psychotherapy, although ratings of adjustment based on TAT protocols showed no change over the period.

Dymond attributes increased self–ideal-self congruence without psychotherapy as due to the strengthening of neurotic defenses. It might be charged that similar changes during therapy might have the same basis. Dymond also raises the possibility that the sorts can be influenced by the attitude of the therapist towards the client's self. There is in short no complete assurance that the cognitive self-acceptance as measured by the *Q* sort is related to the deeper level of self-integration that client-centered therapy seeks to achieve.

Indirect evidence of change of the self-concept during counseling is provided by studies showing changes of self-estimates. Several studies show that agreement between self-ratings on interests and the ratings of the self by interest inventories increase as a result of counseling (Berdie, 1954; Froehlich, 1954; Johnson, 1953; Singer & Stefflre, 1954). The first two of these studies show a moderate increase in accuracy in predicting one's intelligence, but very little improvement in rating the self on measures of personality. One might reason that some parts of the self-concept are peripheral to the core of the self (e.g., interests) and are therefore unstable, while other parts (e.g., personality estimates) are central to the self and are therefore extremely resistant to change.

Self-Concept—Self-Consistency

If the self-concept is to have usefulness as a construct it must be shown that it is consistent in a given self. It must be known whether the self-concept is a gestalt that is more than the sum of different self-regarding attitudes, or whether instead the self-concept is an impossible attempt to generalize different feelings toward unique situations.

One answer to this question is provided by Akeret (1959). He intercorrelated self-ratings on academic values, interpersonal relations, sexual adjustment, and emotional adjustment, achieving differentially positive interrelationships. Emotional adjustment was the best indicator, correlating + .61 with a total corrected for part-whole inflation. While Akeret interpreted his results as suggesting that an individual does not accept or reject himself totally, the results might also be interpreted as suggesting that some areas of self-regard are more central to the self-concept than other areas.

Consistency in the self-concept was found by Martire and Hornberger (1957), who found very great similarities between measures of the actual self, the ideal self, and a socially desirable self. But inconsistency was found by McKenna, Hofstaetter, and O'Connor (1956), who found that one's self ideal differed less from one's close friends than the close friends differed from each other. These investigators concluded by rather involved reasoning that the ideal self is sufficiently differentiated to seek different need satisfactions in different people.

The search for consistency in the self led also to comparing scores on different measures of self-concept. Omwake (1954) compared three scales —the Bills, Phillips, and Berger— which measure acceptance both of self and of others. The scales were in closer agreement as to the degree of acceptance of self than they were as to acceptance of others. Brownfain (1952) found that low ratings of self were related on his scale to the discrepancy between optimistic and pessimistic self-ratings, or what Brownfain termed stability of self-concept; and Cowen (1954) found a relationship between the pessimistic Brownfain self-ratings, and the discrepancy between self– and ideal–self-ratings on the Bills index. Bendig and Hoffman (1957) found that Bills' scores on acceptance of self-ratings and on congruence between ratings of self and ideal self related equally well to scales of the Maudsley Personality Inventory. They therefore concluded that the two different Bills index measures are redundant.

But on the negative side, Cowen (1956) found no relation between the so-called stability of self-concept on the Brownfain, and the different measures on the Bills. Hampton (1955) likewise failed to find any significant relationship between ability to make realistic appraisals about oneself and the ability to admit statements that were damaging but probably true.

Different measures of the self-concept have different theoretical and operational bases. Where measures apply similar rationale, significant correlations between measures have been found. But in similar measures such extraneous variables as response set and social desirability will produce similar bias. Measures of self-concept have reliability, and in a certain degree are interchangeable. Whether or not the reasons for similarity are intrinsic to the scales, the notion of the internal frame of reference seems well validated.

Discussion

The scientist can not hold truths to be self-evident. What is known of the self through direct report must be considered suspect due to philosophical considerations, since the nature of the "I" has been seen differently in each ideological epoch. Notions concerning the self are like other human ideas, and are inventions and not discoveries. The task is not that of discovering the "true self," but instead of constructing those notions which increase understanding of human behavior. Just as the number of inventions is potentially unlimited, so there need be no limit on the number of constructions put upon the self. In this discussion we will proceed functionally, and consider the uses to which different selves have been put.

The first self is the knowing self of structural psychology. Its function is to apprehend reality. The rational nature of man has always been in dispute, and the New Look in perception has further undermined this conception. This article has cited studies which throw doubt on the ability of the self to perceive itself correctly in those areas which are of great value to it. It is the change in the self as perceiver of itself that is the aim of client centered therapy. Studies of client centered therapy do not reveal whether therapy brings the client any closer to reality, but they do provide some evidence that the perception of the self is brought closer to social expectancies.

The second construction of the self is that of motivator. This is the self of thinkers who believe that the individual is motivated by a need for self-assertion, or self-realization, by realizing those potentialities which inhere within the self. Attempts to validate this construct of the self have been carried on through work on *need achievement.* This construct of the self seems involved also in ratings and *Q* sorts for an ideal self which out-distances the real self. Here, of course, the self whose reach exceeds its grasp is considered to be pathological, for it is shown how psychotherapy helps reduce the disparity between the real and ideal.

The third construct of self is the humanistic, semireligious conception of the self as that which experiences itself. It is the "unique personal experience" of Moustakas (1957) and the experience of feeling in Rogers (1951). The difficulty for the psychologist is that such a conception is more religious than scientific; it becomes a value-orientation, and, as the writer has shown elsewhere (Lowe, 1959), it becomes a highly controversial statement of what is the highest good.

The fourth approach views the self as organizer. This self is the psychoanalytic ego; the internal frame of reference of Snygg and Combs (1949); and the source of construct making in G. A. Kelly (1955). Any operational measure of self-consistency would seem to imply the existence of such a self. It is this self that this article has been most directly concerned with; to the extent that studies have been positive, the self does respond the same way in different situations. Conversely, to the extent that the studies have had negative results there is enough inconsistency in the self that it does not always act according to prediction.

A fifth approach constructs the self as a pacifier. Such a self seems implied in Lewin (1936), who constructed his system of personality in terms of valences or tensions which the organism seeks to keep to a minimum. It seems present also in Angyal (1941) who views life as an oscillation about a position of equilibrium. The self in other words is seen as an adjustment mechanism which seeks to maintain congruence between the self and the nonself. It is the verification of this type of self that seems implied by *Q* sort studies that show increased congruence of real and ideal self as a result of psychotherapy. We must however note that the self as pacifier stands in direct opposition to the self as motivater.

In the sixth view of the self, the self is the subjective voice of the culture, being purely a social agent. It is the self of both sociology and S–R psychology, for it sees behavioral responses solely in terms of social conditions or stimuli inputs. The self as an entity is denied, and behavioral consistency is seen as residing not in the individual but in similar environmental events. If the term self is used, it is seen in terms of ego-involvements with loyalties which are determinative of the self.

From these different conceptions of the self, we can choose the one which best fits our theoretical frame of reference. But which conception is chosen seems to depend more upon faith than upon logic, and the choice of one conception must of necessity deny other constructs. It seems impossible that the self can function as a motivator which constantly tries to change the status quo, and as a pacifier which minimizes the disparity between the real and ideal self. There is a contradiction also between the self as motivator and the self as feeling, for in the latter the self is accepted as it is, but in the former is not. Differences are apparent also between the self as feeling and as pacifier. And finally, the self as agent of society is opposed to all other conceptions.

Conclusion

Is the self-concept a fact which, having an objective existence in nature, is observed and measured; or is it an epiphenomenon of deeper reality, invented by man that he might better study his behavior?

The world has sought to be so sure of the self because there is so little else of which it can be certain. The self has become the anchor that man hopes will hold in the ebbtide of social change. But just as a fish could never know it was surrounded by water unless that water were to disappear, it is unlikely that Lecky (1945) would have known about self-consistency had he not lived in a culture which felt inconsistency. In Buberian terminology, the self is an It, which man invents because he can not find a Thou.

The position of this paper must be that the self is an artifact which is invented to explain experience. If the self-concept is a tool, it must be well designed and constructed. We will conclude therefore with that construct of the self which best serves the 1960s. Such a construction combines the self of ego-involvement with the self of feeling. It is a self which is existential not to experience itself, but to mediate encounter between the organism and what is beyond. Such a self is what Pfuetze (1954) calls the "self-other dialogic theory of the self," being interpreted naturalistically through Mead and transcendentally through Buber. It is as an artifact that the self-concept finds meaning.

REFERENCES

AKERET, R. U. Inter-relationships among various dimensions of the self-concept. *J. counsel. Psychol.*, 1959, 6, 199–201.

ALLPORT, G. W. The ego in contemporary psychology. *Psychol. Rev.*, 1943, 50, 451–478.

ANGYAL, A. *Foundations for a science of personality.* New York: Commonwealth, 1941.

BENDIG, A. W., & HOFFMAN, J. L. Bills Index of Adjustment and the Maudsley Personality Inventory. *Psychol. Rep.*, 1957, 3, 507.

BERDIE, R. F. Changes in self-ratings as a method of evaluating counseling. *J. counsel. Psychol.*, 1954, **1**, 49–54.

BERGER, E. M. The relation between expressed acceptance of self and expressed acceptance of others. *J. abnorm. soc. Psychol.*, 1952, **47**, 778–782.

BERTOCCI, P. A. The psychological self, the ego, and personality. *Psychol. Rev.*, 1945, **52**, 91–99.

BILLS, R. E. Rorschach characteristics of persons scoring high and low in acceptance of self. *J. consult. Psychol.*, 1953, **17**, 36–38. (a)

BILLS, R. E. A validation of changes in scores for the Index of Adjustment and Values as measures of changes in emotionality. *J. consult. Psychol.*, 1953, **17**, 135–138. (b)

BILLS, R. E. Self-concepts and Rorschach signs of depression. *J. consult. Psychol.*, 1954, **18**, 135–137.

BILLS, R. E. *Manual for the Index of Adjustment and Values.* Auburn: Alabama Polytechnic Inst., 1958.

BILLS, R. E., Vance, E. L., & MCLEAN, O. S. An index of adjustment and values. *J. consult. Psychol.*, 1951, **15**, 257–261.

BLOCK, J., & THOMAS, H. Is satisfaction with self a measure of adjustment? *J. abnorm. soc. Psychol.*, 1955, **51**, 254–259.

BROWNFAIN, J. J. Stability of the self-concept as a dimension of personality. *J. abnorm. soc. Psychol.*, 1952, **47**, 597–606.

BUTLER, J. M., & HAIGH, G. V. Changes in the relation between self-concepts and ideal concepts consequent upon client-centered counseling. In C. R. Rogers & R. Dymond (Eds.), *Psychotherapy and personality change.* Chicago: Univer. Chicago Press, 1954. Pp. 55–75.

CALVIN, A. D., & HOLTZMAN, W. H. Adjustment and discrepancy between self-concept and inferred self. *J. consult. Psychol.*, 1953, **17**, 39–44.

CARTWRIGHT, R. D. Effects of psychotherapy on consistency. *J. counsel. Psychol.*, 1958, **4**, 15–21.

CHASE, P. H. Self-concept in adjusted and maladjusted hospital patients. *J. consult. Psychol.*, 1957, **21**, 495–497.

CHEIN, I. The awareness of self and the structure of the ego. *Psychol. Rev.*, 1944, **51**, 304–314.

CHODORKOFF, B. Adjustment and the discrepancy between the perceived and ideal self. *J. clin Psychol.*, 1954, **10**, 266–268. (a)

CHODORKOFF, B. Self-perception, perceptual defense, and adjustment. *J. abnorm. soc. Psychol.*, 1954, **49**, 508–512. (b)

COOPERSMITH, S. A method for determining types of self-esteem. *J. abnorm. soc. Psychol.*, 1959, **59**, 87–94.

COWEN, E. L. The "negative self-concept" as a personality measure. *J. consult. Psychol.*, 1954, **18**, 138–142.

COWEN, E. L. Investigation between two measures of self-regarding attitudes. *J. clin. Psychol.*, 1956, **12**, 156–160.

COWEN, E. L., & TONGAS, P. N. The social desirability of trait descriptive terms: Applications to a self-concept inventory. *J. consult. Psychol.*, 1959, **23**, 361–365.

DYMOND, R. F. Adjustment changes over therapy from Thematic Apperception Test ratings. In C. R. Rogers & R. Dymond (Eds.), *Psychotherapy and personality change.* Chicago: Univer. Chicago Press, 1954. Pp. 109–120.

DYMOND, R. F. Adjustment changes in the absence of psychotherapy. *J. consult. Psychol.*, 1955, **19**, 103–107.

EASTMAN, D. Self-acceptance and marital adjustment. *J. consult. Psychol.*, 1958, **22**, 95–99.

EDELSON, M., & JONES, A. E. Operational exploration of the conceptual self-system and of the interaction between frames of reference. *Genet. psychol. Monogr.*, 1954, **50**, 43–140.

EDWARDS, A. L. *Social desirability variables in personality assessment and research.* New York: Dresden, 1957.

ENGEL, M. The stability of the self-concept in adolescence. *J. abnorm. soc. Psychol.*, 1959, **58**, 211–215.

EWING, T. N. Changes in attitude during counseling. *J. counsel. Psychol.*, 1954, **1**, 232–239.

FIEDLER, R. E., DODGE, JOAN S., JONES, R. E., & HUTCHINS, E. B. Interrelations among measures of personality adjustment in non-clinical populations. *J. abnorm. soc. Psychol.*, 1958, **56**, 345–351.

FRIEDMAN, I. Phenomenal, ideal, and projected conceptions of the self. *J. abnorm. soc. Psychol.*, 1955, **51**, 611–615.

FROELICH, C. P. Does test taking change self ratings? *Calif. J. educ. Res.*, 1954, **5**, 166–169.

HAMPTON, B. J. An investigation of personality characteristics associated with self-adequacy. Unpublished doctoral dissertation, New York University, 1955.

HANLON, T. E., HOFSTAETTER, P. R., & O'CONNOR, J. P. Congruence of self and ideal-self in relation to personality adjustment. *J. consult. Psychol.*, 1954, **18**, 215–218.

HILGARD, E. R. Human motives and the concept of the self. *Amer. Psychologist*, 1949, **4**, 374–382.

Holt, R. R. Accuracy of self-evaluations. *J. consult. Psychol.*, 1951, **15**, 95–101.

Johnson, D. G. Effect of vocational counseling on self-knowledge. *Educ. psychol. Measmt.*, 1953, **13**, 330–338.

Kelly, G. A. *Psychology of personal constructs.* New York: Norton, 1955.

Kelman, H. C., & Parloff, M. B. Interrelations among three criteria of improvement in group therapy. *J. abnorm. soc. Psychol.*, 1957, **54**, 281–288.

Lecky, P. *Self-consistency.* New York: Island, 1945.

Lewin, K. *Principles of topological psychology.* New York: McGraw-Hill, 1936.

Lowe, C. M. Value-orientations: An ethical dilemma. *Amer. Psychologist*, 1959, **14**, 687–693.

Lundholm, H. Reflections on the nature of the psychological self. *Psychol. Rev.*, 1940, **47**, 110–127.

McKenna, H. V., Hofstaetter, P. R., & O'Connor, J. P. Concepts of ideal self and of the friend. *J. Pers.*, 1956, **24**, 262–279.

Martire, J. G., & Hornberger, R. H. Self-congruence by sex and between sexes in a "normal" population. *J. clin. Psychol.*, 1957, **13**, 288–291.

Moustakas, C. *The self.* New York: Herper, 1957.

Omwake, K. T. Relation between acceptance of self and acceptance of others shown by three personality inventories. *J. consult. Psychol.*, 1954, **18**, 443–446.

Pfuetze, P. E. *The social self.* New York: Bookman, 1954.

Phillips, E. L. Attitudes toward self and others. *J. consult. Psychol.*, 1951, **15**, 79–81.

Raimy, V. C. The self-concept as a factor in counseling and personality organization. Unpublished doctoral dissertation, Ohio State University, 1943.

Roberts, G. E. A study of the validity of the Index of Adjustment and Values. *J. consult. Psychol.*, 1952, **16**, 302–304.

Rogers, C. R. *Client-centered therapy.* Boston: Houghton Mifflin, 1951.

Rogers, C. R., & Dymond, R. (Eds.) *Psychotherapy and personality change.* Chicago: Univer. Chicago Press, 1954.

Rogers, C. R., et al. A coordinated research in psychotherapy. *J. consult. Psychol.*, 1949, **13**, 149–220.

Rudikoff, E. C. A comparative study of the changes in the concepts of the self, the ordinary person, and the ideal in eight cases. In C. R. Rogers & R. Dymond (Eds.), *Psychotherapy and personality change.* Chicago: Univer. Chicago Press, 1954. Pp. 85–98.

Sheerer, E. J. An analysis of the relationship between acceptance of and respect for self and others. *J. consult. Psychol.*, 1949, **13**, 169–173.

Singer, S. L., & Stefflre, B. Analysis of the self-estimate in the evaluation of counseling. *J. counsel. Psychol.*, 1954, 1, 252–255.

Smith, G. M. Six measures of self-concept discrepancy and instability: Their interrelations, reliability, and relations to other personality measures. *J. consult. Psychol.*, 1958, **22**, 101–112.

Snygg, D., & Combs, A. *Individual behavior.* New York: Harper, 1949.

Stephenson, W. *The study of behavior.* Chicago: Univer. Chicago Press, 1953.

Taylor, C., & Combs, A. Self-acceptance and adjustment. *J. consult. Psychol.*, 1952, **16**, 89–91.

Taylor, D. M. Changes in self-concept with psychotherapy. *J. consult. Psychol.*, 1955, **19**, 205–209.

Turner, R. H., & Vanderlippe, R. H. Self-ideal congruence as an index of adjustment. *J. abnorm. soc. Psychol.*, 1958, **57**, 202–206.

Zimmer, H. Self-acceptance and its relation to conflict. *J. consult. Psychol.*, 1954, **18**, 447–449.

Zuckerman, M., & Manashkin, I. Self-acceptance and psychopathology. *J. consult. Psychol.*, 1957, **21**, 145–148.

(Received May 15, 1960)

VOL. 58, No. 5 SEPTEMBER 1961

Psychological Bulletin

BIOMETRICAL GENETICS AND BEHAVIOR: REANALYSIS OF PUBLISHED DATA[1]

P. L. BROADHURST

Institute of Psychiatry, University of London

AND J. L. JINKS[2]

ARC Unit of Biometrical Genetics, University of Birmingham

It is the purpose of this paper to scrutinize the attempts which have been made to provide quantitative data relating to the inheritance of behavioral characteristics in infrahuman animals, and to reanalyze these data in terms of the polygenic or multifactorial hypothesis of genetical determination. Much of the data derived from the psychological field shows continuous variation and is consequently of the sort which lends itself to such polygenic analyysis, as opposed to that employed in the analysis of discrete characteristics typical of classical Mendelian genetics. While it should be noted that there are now several experimental methods and analyses which have been developed for dealing with polygenic inheritance, it is not our present intention to undertake an evaluative survey of their relative merits as applied to psychogenetics at its current stage of development. Instead we propose to concentrate on and to employ one set of such techniques, those of biometrical genetics as developed by Mather (1949) expressly for the analysis of continuous variation, especially in plants, and which we judge to be particularly promising in their application to the inheritance of behavior. An introduction to the general model and assumptions of this biometrical approach as applied to psychogenetics will be found in Broadhurst (1960).

EXPERIMENTAL METHOD

Few experiments in psychogenetics have been of a kind which can lead to a partitioning of the variation into its heritable and nonheritable components. Even fewer have been designed in such a way that the various tests are sensitive and the analysis reliable. Satisfactory experimental procedures for applying biometrical analysis in psychogenetics have recently been discussed by Broadhurst (1960) and we will merely note the following points: (*a*) the experimental animals should be randomized and the experiments replicated; (*b*) the parental stocks must be inbred; and (*c*) for investigating a cross between two inbred strains at least the two parental (P), first and second filial (F_1 and F_2) and backcross (B) generations should be reared.

[1] A summarized version of this paper was read to the Society for Experimental Biology in London in January 1961.

[2] Harkness Fellow of the Commonwealth Fund in the Division of Biology, California Institute of Technology, Pasadena, California, 1959-60.

The published data which come nearest to satisfying these requirements are those of Dawson (1932), Brody (1942), Goy and Jakway (1959), Jakway (1959), and Thompson and Fuller.[3] Although no special precautions were taken by Dawson and by Brody to ensure the homozygosity of the parental strains the genetical differences between them were much greater than those within them; and insofar as our interest is in the differences between the parental strains, differences within them may be regarded as a further source of error variations, like the nonheritable differences with which they are confounded. The procedures appropriate to the analysis of these five sets of data are outlined in the next sections.

Analysis of Means

Following Hayman and Mather's (1955) and Jinks and Jones' (1958) extension of Mather (1949) we can write the generation means of a cross between two inbred strains in terms of six parameters: m, [d], [h], [i], [j], and [l] which are the mean, additive, dominance, and three nonallelic, first-order interaction components between pairs of genes, respectively.

Thus:

$$\overline{P}_1 = m + [d] + [i] - \tfrac{1}{2}[j] + \tfrac{1}{4}[l]$$

$$\overline{P}_2 = m - [d] + [i] + \tfrac{1}{2}[j] + \tfrac{1}{4}[l]$$

$$\overline{F}_1 = m + [h] + \tfrac{1}{4}[l]$$

$$\overline{F}_2 = m + \tfrac{1}{2}[h]$$

$$\overline{B}_1 = m + \tfrac{1}{2}[d] + \tfrac{1}{2}[h] + \tfrac{1}{4}[i]$$

$$\overline{B}_2 = m - \tfrac{1}{2}[d] + \tfrac{1}{2}[h] + \tfrac{1}{4}[i]$$

Hence from these generation means we can estimate the heritable components as:

$$[d] = \overline{B}_1 - \overline{B}_2$$

$$[h] = \overline{F}_1 - 4\overline{F}_2 - \tfrac{1}{2}\overline{P}_1 - \tfrac{1}{2}\overline{P}_2 + 2B_1 + 2B_2$$

$$[i] = 2\overline{B}_1 + 2\overline{B}_2 - 4\overline{F}_2$$

$$[j] = 2\overline{B}_1 - \overline{P}_1 - 2\overline{B}_2 + \overline{P}_2$$

$$[l] = \overline{P}_1 + \overline{P}_2 + 2\overline{F}_1 + 4\overline{F}_2 - 4\overline{B}_1 - 4\overline{B}_2$$

and their sampling errors (SE^2) as:

$$V_{[d]} = V_{\overline{B}_1} + V_{\overline{B}_2}$$

$$V_{[h]} = V_{\overline{F}_1} + 16V_{\overline{F}_2} + \tfrac{1}{4}V_{\overline{P}_1} + \tfrac{1}{4}V_{\overline{P}_2} + 4V_{\overline{B}_1} + 4V_{\overline{B}_2}$$

$$V_{[i]} = 4V_{\overline{B}_1} + 4V_{\overline{B}_2} + 16V_{\overline{F}_2}$$

$$V_{[j]} = 4V_{\overline{B}_1} + V_{\overline{P}_1} + 4V_{\overline{B}_2} + V_{\overline{P}_2}$$

$$V_{[l]} = V_{\overline{P}_1} + V_{\overline{P}_2} + 4V_{\overline{F}_1} + 16V_{\overline{F}_2} + 16V_{\overline{B}_1} + 16V_{\overline{B}_2}$$

The standard errors of the components can thus be obtained and tests of their significance by the customary methods applied.

If the gene effects are additive, that is, the genes are independent in their action, then the three components which estimate the effects of nonallelic interactions, [i], [j], and [l], will be nonsignificant and the following identities known as *scaling tests* will hold within the limits of sampling error (Mather, 1949):

$$\text{Test A:}\quad \overline{P}_1 + \overline{F}_1 - 2\overline{B}_1 = 0$$

$$\text{Test B:}\quad \overline{P}_2 + \overline{F}_1 - 2\overline{B}_2 = 0$$

$$\text{Test C:}\quad \overline{P}_1 + \overline{P}_2 + 2\overline{F}_1 - 4\overline{F}_2 = 0$$

A joint test of these three identities has been devised by Cavalli (1952). In this we estimate weighted least squares values for m, [d], and [h] from the generation means, assuming the absence of nonallelic interactions. The weights used are the reciprocals of the squared standard deviations of the generation means. The squared deviations of the ex-

[3] W. R. Thompson and J. L. Fuller, personal communication, 1959.

pected and observed generation means are then a χ^2 with $(n-3)$ degrees of freedom, where n is the number of observed generation means.

If this χ^2 is nonsignificant then nonallelic interactions are absent and we can interpret directly the estimates of [d] and [h] obtained in the scaling test. The ratio of

$$\frac{[h]}{[d]} = \frac{h}{r_d \Sigma d}$$

is the so-called "potence ratio" (Wigan, 1944) and this is a measure of dominance only if the genes are associated in the parent lines, that is r_d (the degree of association) $= \pm 1$ (Jinks & Jones, 1958) and all [h] increments have the same sign. The potence ratio can theoretically take any value between zero and infinity. While a significant potence ratio indicates dominance of the individual genes predominantly in the same direction, zero potence does not necessarily indicate absence of dominance.

If the χ^2 from the joint scaling test is significant then nonallelic interactions are present and these can be analyzed by estimating [i], [j], and [l] and testing their significance. A comparison of the signs of [l] and [h] will then tell us the type of nonallelic interaction involved. If their signs are the same then cooperative or complementary interaction between the genes predominates while if their signs differ competitive or duplicate interaction predominates (Jinks & Jones, 1958). The component $[j] = r_j \Sigma j$ on the other hand provides us with an indication of the distribution of the interacting genes in the parental lines. Thus with complete association $r_j = \pm 1$ and [j] may have a significant value, but with maximum dispersion $r_j = 0$ and [j] must be zero.

Scales

Although allowance can be made for nonallelic interactions and genotype-environmental interactions in the analysis of second degree statistics (Hayman & Mather, 1955; Mather & Jones, 1958), with the paucity of second degree statistics available in the psychogenetical experiments to be analyzed we can merely attempt to find an empirical scale on which these effects make no significant contribution to the variation.

Clearly, our criterion of an adequate scale which eliminates nonallelic interaction is one which leads to a nonsignificant χ^2 in the joint scaling test. However, a scale which is empirically adequate for this purpose will not necessarily remove any genotype-environmental interaction which may be present, that is, lead to homogeneity of the variances of the parents and F_1s. We must, therefore, adopt a scale which at least minimizes and balances these two sources of bias.

Analysis of Variances into Components of Variation

On an adequate scale the variances (s^2) of the parent, F_1, F_2, and backcross generations are (Mather, 1949):

$$V_{P_1} = V_{P_2} = V_{F_1} = E_1$$

$$V_{F_2} = \tfrac{1}{2}D + \tfrac{1}{4}H + E_1$$

$$V_{B_1} + V_{B_2} = \tfrac{1}{2}D + \tfrac{1}{2}H + 2E_1$$

where $D = \Sigma d^2$, $H = \Sigma h^2$ and E_1 are the additive, dominance, and nonheritable components of variation, respectively.

In addition we have

$$V_{B_1} - V_{B_2} = \pm \Sigma(dh)$$

Solution of these equations leads to estimates of D, H, and E_1, from which we can obtain estimates of dominance and heritability. The dominance ratio, H/D, will be zero for no dominance, one for complete dominance and greater than one for overdominance (heterosis). Heritability can be assessed in a variety of ways of which we will use two, $D/(D+E_1)$, i.e., the ratio of the additive variation to the sum of the additive plus nonheritable variation, and $(\frac{1}{2}D+\frac{1}{4}H)/(\frac{1}{2}D+\frac{1}{4}H+E_1)$, which is the proportion of heritable variation in an F_2 population. These ratios therefore represent estimates of heritability "in the narrow sense" and "in the broad sense," respectively.

When $\Sigma(dh)$ does not equal zero it supplies additional evidence for the presence of dominance. It also shows which parent carries the preponderance of dominant allelomorphs, for the backcross to this parent has the lower variance.

Number of Effective Factors

Only one estimate of the number of effective factors is applicable to the type of data so far obtained in psychogenetics, namely, the estimate of K, (Mather, 1949). This equals $\frac{1}{4}(P_1-P_2)^2/D$ which for k genes of equal effect and associated in the parental lines equals k^2d^2/kd^2.

In practice this estimate is always minimal because it assumes that the genes are associated (i.e., $r_d=1$) and that all genes give equal increments (i.e., $d_a=d_b=d_c\cdots$). It is, however, worth obtaining in the psychogenetical experiments because of the practice of deliberately selecting the most extreme lines available as parents in the cross. Such selection will lead to a preponderance of association in the parental lines, thus partially satisfying one of the assumptions required.

Examples from the Literature

We can now illustrate the analyses described earlier by reference to particular experiments in the field of psychogenetics.

Dawson

Dawson's work (1932) was most accomplished genetically, and still might serve as a model of how a psychogenetical investigation could be approached, at least, from the genetical aspect. Unfortunately, the more purely psychological treatment is not of comparable quality and leaves much to be desired. Dawson investigated the inheritance of wildness in mice, defining wildness in terms of the speed the animals showed in running down a straight runway. We shall give his description of the method.

> The method of testing consisted in placing the mouse at one end of a runway and allowing it to run to the other end. The time required was recorded by means of a stopwatch. The runway was 24 feet long, 9½ inches wide and 13 inches high. The sides and ends were of galvanized sheet iron, the floor of soft wood. One foot from each end a black line was painted on the floor of the runway. The time required for the mouse to run from one line to the other, a distance of 22 feet, was recorded. A movable partition made of wallboard and bound with rubber was used to prevent the mouse from running back during the test and to aid in starting the test and capturing the mouse afterwards. The following procedure was carried out in testing the mice in this device. On the date that the mice in a certain pen were to be tested they were carried to the runway a short distance away and tested one at a time. The mouse to be tested was confined by the movable partition in a space about one foot from the end of the runway until everything was ready when the partition was raised and the stopwatch started as soon as the mouse crossed the black line. The mouse was followed by the experimenter with the partition which was placed in position to prevent the animal from running back if it showed any signs of doing so. Nothing was done to frighten the mouse other than the procedure described. This was usually

> sufficient to cause even the tame mice to run or walk towards the other end of the runway. If the mouse ran swiftly, it was impossible to keep up with it with the partition; but if more slowly the partition was moved along and kept about twelve to fifteen inches behind the mouse. If the mouse stopped and showed no inclination to go forward the partition was slowly advanced until it touched the mouse. In all but three or four cases this was sufficient to start the mouse again. The few individuals where this was not the case were shoved a little and thus started. When the mouse crossed the line at the far end of the runway, the watch was stopped and the partition taken out of the runway allowing the mouse to run back to the starting point or in case it did not do so voluntarily it was urged by means of the partition. This prevented the mice from associating the far end of the runway with being caught. Since each individual was tested three times this point was of considerable importance. After the mouse had been cornered at the starting point by means of the partition, it was caught and the number in its ear read. . . . These trials were conducted at weekly intervals after the mouse reached 75 days of age. In order to facilitate the testing and caring for the mice, a variation of one day in either direction was permitted. Thus the first trial for a given mouse might occur on the 74th, 75th or 76th day. A few trials had to be made on different dates. The trials were in nearly all cases made in the evening or at night when there was very little outside disturbance to distract the mice. The lighting was kept as far as possible the same throughout the experiment (pp. 299–300).

It will be seen that a large subjective element could enter into the factors determining the speed of running of a given subject in the alley through the way in which the partition was manipulated. This is not entirely overcome by the procedure of identifying the animal after completing the test, as, in the parental generations at any rate, there were distinctive coat color differences between the strains. Nevertheless, the corrected reliability coefficient for the three trials for all the 1,232 subjects used is reported as 0.92 ±(SE)0.04.

The subjects were a strain of wild mice which had been reared in the laboratory for several years and three strains of tame mice, an albino, and two strains of brown mice with pink eyes, one of them also having short ears. The wild strain of mice appeared to be more highly inbred than the tame mice since selection among the former produced no response while the latter responded. This, however, as we have seen earlier, is not a serious problem in view of the large and high significant differences between the wild and tame strains for the measure under consideration.

Dawson was able to extract a large amount of information from his results regarding the nature of the genetical control of behavior in his runway situation. He showed that there was no linkage with sex or with any of the major gene effects identifiable in his strains, and, by reciprocal crossing, that there were probably no directional maternal effects (Broadhurst, 1961). He concluded that the wild-type behavior was dominant and that only a few genes are involved in determining the reaction. This last conclusion is principally based on the result of fitting curves, derived from Mendelian ratios and assuming various numbers of genes up to three, to the observed distributions. However, he admitted that probably a number of modifying genes were also involved. Implicit in his estimate of the number of genes were assumptions concerning size of individual gene effects, the distribution of genes in the parental lines and their dominance relations. A biometrical analysis along the lines proposed here would, therefore, appear to be appropriate.

In Table 1 will be found the relevant generation means and their standard errors calculated from the data given by Dawson. The three

TABLE 1

DAWSON'S DATA: MEANS AND THEIR STANDARD ERRORS IN SECONDS AND (*n*)

	Generation Means						
	P_1 (Wild)	P_2 (Tame)	F_1	F_2	B_1	B_2	
Males	6.7±0.3 (43)	24.5±1.0 (63)	7.6±0.3 (76)	13.0±0.6 (175)	6.6±0.3 (26)	27.4±3.9 (54)	20.8±1.6 (50)
Females	5.3±0.3 (47)	25.3±1.2 (54)	6.9±0.3 (88)	11.8±0.5 (190)	6.2±0.5 (24)	18.7±1.5 (48)	
Both	5.9±0.2 (90)	24.9±0.8 (117)	7.2±0.2 (164)	12.4±0.4 (365)	6.4±0.4 (50)	23.3±2.2 (102)	19.7±1.4 (98)

strains of tame mice were used in these crosses and the results pooled.

Apart from the backcross to the slower parent (B_2), the sexes are in good agreement and this failure in the backcross can be traced to four males whose scores were greater than 90. Individuals with such high scores are met with nowhere else in Dawson's experiments which included some second backcrosses, F_3s and F_4s in addition to the data given in the table. The results omitting the four males are also shown in Table 1. Their omission improves the already good agreement between sexes and the analyses can now be carried out on the pooled sexes. We may add that omitting these four individuals does not affect the interpretation of the data since the sex difference they suggest is borne out neither by Dawson's nor our own detailed analyses.

The joint scaling test (Cavalli, 1952) gives the following weighted least squares estimates from the pooled sexes,

$$m = 15.99,\ [d] = 10.10 \text{ and } [h] = -8.74$$

which when compared with the observed generation means give a $\chi_{(3)}^2$ of 9.4 ($p = 0.05 - 0.02$). There is therefore some nonallelic interaction present. However, its magnitude would not normally warrant rescaling but these data also show significant geno-type-environmental interaction ($p < 0.01$) on the linear scale, as determined by Pearson and Hartley's test (1958) for inhomogeneity of the P_1, P_2, and F_1 variances. Two transformations have, therefore, been tried, a square root and a log transformation. Of these the latter was the more satisfactory and the joint scaling test repeated on the new scale gave $m = 1.197 \pm 0.032$, $[d] = 0.340 \pm 0.030$, and $[h] = -0.222 \pm 0.059$ which gave a satisfactory fit with the observed data ($\chi_{(3)}^2 = 0.34$).

The same scalar change also removed the genotype-environmental interaction; hence a solution for D, H, and E_1 was attempted from the second degree statistics. These gave values[4] of

$$D = 0.052 \pm 0.024$$

$$H = -0.008 \pm 0.032$$

$$E_1 = 0.020 \pm 0.005$$

$$\Sigma(dh) = -0.032$$

This gives a heritability estimated as $D/(D+E)$ of 72% and estimated as $(\frac{1}{2}D + \frac{1}{4}H)/(\frac{1}{2}D + \frac{1}{4}H + E)$ of 55%. The confidence limits ($p = .05$) for the first estimate of heritability are 61% and 83%. A further estimate of heritability can be extracted from Dawson's data. He assortatively mated his F_2 individuals to raise an F_3 generation and from the results

[4] This is the only place in this paper where the data permit the estimate of standard errors for these components. No further suitable replication of observations, e.g., by the provision of the raw data for both sexes, as in this case, is encountered.

we can estimate the parent/offspring correlation. This turns out to be 0.51.

Our estimate of the minimal number of effective factors is 2.2 with confidence limits of 3.5 and 1.6. This is in good agreement with Dawson's estimate which as we have seen is also minimal.

Thus, the behavioral difference between the wild and tame mice investigated by Dawson is controlled by at least three effective factors whose contributions are additive and independent of the environment on a logarithmic scale but which interact with one another and with the environment on a linear scale. Estimates of [i], [j], and [l] on the linear scale show that [j] which equals 7.6 ± 2.9 is responsible for the nonallelic interactions. The genes have a significant additive and dominance effect although the latter is not apparent in the second degree statistics presumably due to the effect of sampling variation on the negative correlation between the estimates of D, H, and E_1. The potence ratio is negative and greater than zero which means that there is a preponderance. of dominant genes in the low scoring, i.e., wild type, parent. The significant estimate of $\Sigma(dh)$ confirms this and also shows the presence of a dominance component of variation. Heritability is quite high and estimates from different sources give consistent results.[5]

Brody

The analysis of Brody's experiments (1942) follows much the same pattern. She investigated the inheritance of voluntary cage activity in rats using the high and low selections for activity begun by Rundquist (1933). The number of revolutions of the activity cages, during the last 15 of a 21-day period was taken as the measure. Each rat was housed in these cages at some time between 60 and 100 days of age. Some measure of inbreeding was practiced from the fifth generation of selection although neither its degree nor the precautions taken to control environmental variation are stated. Our present concern is with the crosses made using the inactive and active strains at the twenty-first generation of selection. A complete program of breeding F_1, F_2, and backcrosses, B_1 and B_2, was carried out, and, moreover, repeated using these two strains at the twenty-second generation of selection as parents. In each case the strains were crossed reciprocally to give the F_1s.

Brody's results show reasonably good agreement between the values obtained in the two replications of her crossing program. An analysis of variance of the complete data gave significance for only two items: the difference between sexes, and the difference between generations, i.e., P_1, P_2, F_1, F_2, B_1 and B_2. There were no significant differences between the replicate crossing programs initiated at the twenty-first and the twenty-second generations of selection and no significant interactions between the three main effects. We can, therefore, pool the two sets of crosses and sexes for the biometrical analyses.

The joint scaling test on the pooled data given in Table 2 gave weighted least squares estimates of $m = 74.71 \pm 5.55$, $[d] = 63.03 \pm 5.56$ and $[h] = -3.41 \pm 9.32$. These did not pro-

[5] Since this paper was submitted for publication, estimates of the number of genes and heritability in the F_2, derived from another reanalysis of Dawson's data, have been published by Fuller and Thompson (1960). They used methods of analysis (Wright, 1952) similar to those proposed here, and their estimates are in substantial agreement with our own.

TABLE 2

BRODY'S DATA: MEANS AND THEIR STANDARD ERRORS IN REVOLUTIONS × 10^{-3} AND (n)

Pooled Generation Means					
P_1 (Inactive)	P_2 (Active)	F_1	F_2	B_1	B_2
13.2±4.1 (136)	136.4±12.5 (67)	73.6±7.1 (193)	71.5±9.3 (260)	26.8±9.3 (79)	115.2±10.2 (136)

vide a satisfactory fit with the observed generation means, $\chi_{(3)}^2=15.27$ ($p=0.01-0.001$). Clearly genic interaction is present on this scale. Since these data are not published in a form amenable to rescaling we cannot attempt to remove the interaction; we can, however, investigate its nature.

Estimates of the components of the generation means and their standard errors showed that only m, [d], and [j] were significant hence weighted least squares estimates of these three components were made assuming the other components were zero. These estimates of $m=72.14\pm3.61$, $[d]=81.54\pm13.62$ and $[j]=44.34\pm28.90$ provided a satisfactory fit with the observed generation means, $\chi_{(3)}^2=5.28$ ($p=0.20$–0.10).

A further argument in favor of rescaling is provided by the second degree statistics which show significant ($p=0.01-0.001$) genotype-environmental interaction. Since, however, rescaling is impossible we must be cautious in interpreting the components of variation because of the possible bias from the [j]-type nonallelic interaction and the genotype-environmental interactions. H proved to be nonsignificant and negative therefore D and E were recalculated assuming $H=0$ with the following result: $D=3410.72$ and $E=1945.91$.[6] These values give 64% and 44% as our two estimates of heritability and 1.95 as the minimal number of effective factors. The uncertainty of the latter estimate makes it impossible decisively to rule out Brody's own interpretation based on a single gene difference between her selected parents. On the other hand we can discount a single gene interpretation on the basis of the significant genic interaction.

Thus the difference between the spontaneous cage activity of the selected strains measured on a linear scale depends on at least two interacting genes which also interact with the environment. As in Dawson's experiment it is the [j]-type interaction which is responsible for the genic interaction. There is no evidence of dominance and the potence ratio is zero. It is possible, however, that dominant and recessive alleles are equally frequent in the two selected strains. Heritability is about the same as in Dawson's experiment but since the scale on which it is measured is unsatisfactory we cannot place too much reliance on its absolute magnitude.

An unusual feature of Brody's results requires further comment. Both her F_1s were made reciprocally. Both of them show the same paternal effect, that is, in the direction of a *negative* influence of the mother—mothers from the active strain tending

[6] Despite Brody's replication, no estimate of standard errors can be given for these components which is not potentially subject to serious inflation due to the various differences between the variances noted. The impossibility of rescaling therefore renders the replicates unsatisfactory as a source of an estimate of error variation.

to have offspring lower in activity than those from inactive mothers. Pearson and Hartley's (1958) exact test for homogeneity of variance showed only the data from the twenty-first generation crosses were suitable for analysis of variance, which was applied following Snedecor's method (1956) for dealing with unequal numbers in subgroups in a two-way classification. This analysis revealed an interaction between sex and strain as shown in Table 3, which may be summarized by saying that the significant tendency of the F_1 to be unlike the mother in respect of activity was more pronounced in the case of sons than of daughters. Sex linkage or paternal inheritance may be responsible for this complex situation, but to distinguish between the two would require a much more involved experimental design than that used by Brody.

TABLE 3

BRODY'S DATA: MEANS OF RECIPROCAL CROSSES IN REVOLUTIONS $\times 10^{-3}$ AND (n)

	F_1 Offspring	
Strain of Mother	Males	Females
Active	33.9 (7)	109.4 (24)
Inactive	87.9 (21)	120.7 (30)

Thompson and Fuller

The systematic program of research in psychogenetics which has been proceeding at the Roscoe B. Jackson Memorial Laboratory at Bar Harbor, Maine, for the last decade, has, as might be expected, produced work of high quality. Only one set of data, however, has so far become available[7] which lends itself to the complete biometrical analysis proposed in this paper. This is the work of Thompson and Fuller (Fuller & Thompson, 1960, pp. 267–269; Thompson 1953, 1956, see Footnote 3).

Thompson and Fuller employed the two inbred strains of mice showing extremes of high and low activity from a previous study in which a total of 15 strains had been studied, and tested a large number of subjects on each of two tests which they describe as follows:

> [Test 1] consisted of an open-field 30 by 30 inches with walls $3\frac{3}{4}$ inches high, and a hinged wire-mesh top. The floor was painted gray and the walls a flat black. The floor was divided by lines into a grid of 36 quares, each 5 by 5 inches. At the base of every other square was placed a barrier, 5 by $3\frac{3}{4}$ by $3\frac{3}{4}$ inches, painted a flat black. Leading into the open-field at one corner was a starting box with a separate hinged top. Test 2 was a Y-maze with arms $11\frac{1}{2}$ inches long by 3 inches wide by $3\frac{1}{2}$ inches deep. Angles between arms were equal. One arm was painted black inside, another gray and the third white. The maze was covered by a removable wire-mesh top. An animal was started at the end of the gray arm farthest from the junction point. Observation of animals in both tests were made under dim illumination as follows: in Test 1, a record was made by a mechanical counter of the number of lines crossed by a mouse in a 10 minute period. In Test 2, a count was made of the number of half-arm units traversed during each of six 100 second periods. . . . The correlation between the two tests was approximately 0.60 (Thompson & Fuller, see Footnote 3).

The two parental strains are known to be highly inbred, but the measures taken to control environmental variation are not specified. The F_1, F_2, and backcrosses were bred from the two strains and given both tests. The possibility of order effects of one test upon the other in the resulting data is not discussed. Reciprocal crosses were made, and the results pooled,

[7] We are indebted to W. R. Thompson for making a draft copy of the MS containing these data available to us prior to publication.

as were those of the two sexes as shown in Table 4. Thompson and Fuller subjected the data from the first test to a square root transformation in order to equalize the variance of the parental and F_1 generations which it did successfully. The transformed data is included in Table 4.

The joint scaling test for Test 1, Test 2, and the transformed Test 1 data gave the following weighted least squares estimates:

Component	Test 1	Test 1 (transformed)	Test 2
m	267.6±10.9	12.50±0.41	166.1±4.1
[d]	257.8±10.5	9.98±0.39	82.6±4.0
[h]	29.2±20.5	5.13±0.76	−2.0±7.7
$\chi^2_{(3)}$	24.3 ($p<0.001$)	6.6 ($p=0.10$–0.05)	16.9 ($p<0.001$)

On the linear scales both Test 1 and Test 2 show unsatisfactory fits with additive gene action. Hence we have nonallelic interactions present. Analysis of these nonallelic interactions shows that the [j]-type interaction is again largely responsible for the failure of the linear scale, having values of 25.6±8.3 in Test 1 and 55.8±9.0 in Test 2. Unfortunately the data given by Thompson and Fuller do not suffice to allow rescaling. Analysis of their square root transformed data for Test 1 has, however, shown that rescaling can reduce the genic interaction. Since this scale also removes the significant genotype-environmental interaction in Test 1 we can confidently analyze the second degree statistics on the square root scale.

Only Test 1 on the linear scale, owever, provides estimates of the omponents of variation D, H, and E_1 which are sensible, that is to say, ive positive values for D: the other vo tests do not. This result is not nexpected for Test 2 where interactions are present but it is unexected for the square root transrmed data of Test 1 where the interactions are largely scaled out. We will, therefore, merely give estimates of the heritabilities for the F_2 population since this does not require the partitioning of the heritable components of variation (see above). For Test 1 the values are 73% and 53% for the linear and square root scales, respectively, and for Test 2, 26%. The low value in the latter test provides an additional reason for the failure to obtain sensible estimates of the additive and dominance components of variation. With no estimate of D we cannot evaluate our estimate of the number of effective factors.

Thus the genes controlling the two behavior patterns in mice investigated by Thompson and Fuller have a large additive effect but show no dominance on the linear scale. They do, however, interact with one another and with the environment in a way which can be largely removed by a square root transformation. Once again it is the [j]-type of interaction which is mainly responsible for the failure of the linear scale. On the square root scale there is a preponderance of dominance for higher activity in Test 1. The above conclusions are based on the analysis of means: in this case the failure of the analysis of the components of variation merely serves to confirm the presence of interaction.

Jakway and Goy

A fourth set of data susceptible to a complete biometrical analysis has recently become available. It relates to the analysis of sexual behavior in the male and female guinea pig

(Goy & Jakway, 1959; Jakway, 1959). Two highly inbred strains, whose history is documented as far back as 1906, and whose near homozygosity was established as early as 1927 by the method of exchanging tissue grafts (Loeb & Wright, 1927), were crossed, and F_1s, F_2s, and both backcrosses were bred, reciprocally in each case.

In the study of the inheritance of sexual behavior in *female* guinea pigs, the response to the injection of a controlled amount of female hormones in previously ovariectomized subjects was assesssed in terms of four behavioral measures. No details are given of precautions taken to minimize environmental effects. The test technique is described as follows:

> The median age at the time of ovariectomy was 3.5 months in each genetic group. The distributions were not skewed. . . . Tests of reproductive performance began one month later on the average. For the first 3 tests, each animal was injected with 100 I.U. of oestradiol benzoate followed 36 hours later with 0.2 I.U. of progesterone. . . . The volume of all injections was constant (0.5 c.c.), and injections were given subcutaneously in the left axilla. Immediately after injection with progesterone, the animals were placed in a standard observation cage (in groups of 6 to 12 individuals) and observed continuously for 14 hours. Each animal was tested once every hour to determine the time of appearance of the lordosis reflex. . . . The first lordosis obtained was regarded as the onset of oestrus, and animals failing to respond on any of the 14 hourly tests were viewed as not in oestrus. For those animals responding on at least one hourly test, oestrus was regarded as terminated when they failed to lordose on two successive hourly tests (Goy & Jakway, 1959, pp. 142–143).

The measures used are detailed in another paper (Goy & Young, 1957) as follows:

> (1) Latency of heat is the length of the interval between the injection of progesterone and the elicitation of the first lordosis. (2) Duration of heat is the number of hours lordosis can be elicited. Toward the end of a heat period, lordoses become feeble and difficult to

TABLE 4

THOMPSON AND FULLER'S DATA: MEANS AND THEIR STANDARD ERRORS IN UNITS AS INDICATED AND (n)

Measure	Generation Means					
	P_1 (C57BR)	P_2 (A/Jax)	F_1	F_2	B_1	B_2
Test 1. Number of lines crossed	532.2±23.6 (40)	11.4±5.2 (40)	302.9±16.6 (38)	287.9±16.0 (92)	148.3±14.3 (58)	395.9±18.8 (63)
Test 1. Square root transformation	22.9±0.5	1.9±0.6	17.0±0.6	16.1±0.6	11.0±0.7	18.9±0.7
Test 2. Number of half maze units traversed	248.8±11.1	79.6±6.5	160.6±5.4	164.9±4.4	131.5±3.7	206.2±2.7

> elicit, and an operational criterion is necessary to determine when an animal shall be classified as unresponsive. For this purpose, each animal is stroked or fingered five times and if no lordosis is displayed the animal is considered to be out of heat. . . . (3) The duration of maximum lordosis in seconds. . . . The lordosis reflex includes several components, an arching or straightening of the back, elevation of the pudendum, displacement of the rear feet laterally and caudally so that a wide stance is taken, and emission of a low gutteral growl. When an estrous female is stroked lightly in a caudo-cephalad direction all components of the lordosis are displayed nearly simultaneously. If the stroking is continued (prolonged stimulation), the full reflex will be maintained for a time which varies with the genetic background and the phase of estrus. If stimulation is continued until voluntary termination is produced, the duration of the reflex can be measured with a stop-watch. The response may be considered terminated when any one of the following signs is evident: (a) a sudden or gradual return of the back and pudendum to a normal position; (b) a sudden return of the feet to the normal position and a loss of the wide stance characteristic of the reflex; (c) kicking with the hind feet; (d) dashing forward; (e) squatting; (f) urinating; and (g) an abrupt termination of the growl accompanied by a soft squeal. A stop-watch is started immediately on display of lordosis and stopped as soon as the complete response is no longer apparent. . . . (4) Male-like mounting behavior. Mounts accomplished by an individual are classified as (a) complete mounts at the posterior end including pelvic thrusts, (b) posterior mounts without pelvic thrusts, and (c) abortive mounts which are not posteriorly oriented, do not involve clasping, and usually are not accompanied by pelvic thrusts. Recorded mounting activity is usually preceded by locomotor activity best described as prowling or standing in one place and treading the floor of the cage with the hind feet. Both treading and prowling, when they precede mounting, are accompanied by the typical low gutteral growl or chatter. (5) Per cent of females brought into heat by the hormonal treatment (pp. 342–343).

The means, together with SEs for the first four measures, are given in Goy and Jakway's Table 1,[8] and are not repeated here.

> The data . . . were not normally distributed and the variances of the different genetic groups were unequal. Because of these characteristics, conventional parametric analysis was not feasible. Therefore only . . . non-parametric statistics were employed in the analysis (Goy & Jakway, 1959, p. 143).

However, in the case of Measures 2, 3, and 4 above, distributions in the form of proportions are given which has enabled rescaling as necessary.

The methods used with the *male* guinea pigs in rearing, testing, and scoring their sexual behavior are described by Jakway (1959) as follows:

> The animals were left with their own dams and siblings until weaning on day 25. They were then placed in individual cages 2 ft. × 2 ft. × 1 ft. with two females of their own age. The caging in such groups assured each male of the contact with other animals which is necessary to bring out the behavioural differences between males of the two inbred strains. On day 73 the female cagemates were removed. Between the ages of 77 and 120 days each animal was observed in seven, approximately weekly, 10-minute tests with oestrus females. The mean score from this number of observations is expressive of the mating performance of a given animal. Elements or measures of sexual behaviour . . . are defined as follows: *Circling* is the term employed when the male circles the female. *Sniffing and nibbling* is recorded each time the nose of the male touches the female other than in the anogenital region. *Nuzzling* is recorded when the nose touches the anogenital region of the female. *Mounting* is scored when the male places both forepaws on the female. *Intromission* is recorded when the penis penetrates the vaginal orifice. This is accompanied by rhythmic pelvic thrusts. *Ejaculation* is accompanied by a convulsive contraction of the haunches and terminates the display of sexual behaviour. A test score is a numerical value reflecting three factors: the interval of ejaculation (latency of ejaculation), the amount of sexually oriented activity, and the maturity level of the behaviour. With the exception of circling which is not scored, each measure is given a numerical value from the

[8] The standard errors for the measure "number of mounts per oestrons" were recalculated from the distributions given in Goy and Jakway's Table 3. Our values are in general agreement with those given in their Table 1, taking account of distortion due to grouping, with the exception of Strain 13. The obviously erroneous value given in the last column of their Table 3 may be responsible for the discrepancy.

lowest for sniffing and nibbling to the highest for ejaculation. The value of each is then multiplied by a factor expressive of latency of ejaculation; the shorter the latency, the higher the factor. Since most tests in which ejaculation occurred were terminated before the end of the tenth minute, measurements other than scores will be expressed as rates/15 seconds. Inasmuch as a sexual behaviour score can be attained in several ways, the components were analysed separately for possible patterns of inheritance (p. 151).

The means and standard errors for each component and the composite score are grouped together in Table 5. In each case Jakway gives the percentage distributions which has enabled us to rescale the data as necessary.

The results of the joint scaling tests are summarized in Table 6. Only two measures, *duration of maximum lordosis* in females and *number of ejaculations* in the males show genic interactions on the chosen scales. All the measures show significant heritable variation and only one, *circling* in males shows no significant dominance.

For the four female measures the potence ratio [h]/[d] is approximately half and for three of the measures, *latency of estrus*, *duration of maximum lordosis*, and *frequency of mounting* it is also negative. That is, for these measures the parent with the lower score contains a preponderance of dominant genes. For the male measures the potence ratios are more variable, ranging from non-significant for *circling* to greater than 10 for *number of ejaculations;* in fact for three measures, *nuzzling*, *intromissions*, *ejaculations*, as well as for the composite score, the ratio is greater than one, that is, all these measures show heterosis.

Estimates of the components of the generation means and their standard errors showed that the [j]-type interaction is primarily responsible for the significant deviation for additivity in *duration of maximum lordosis* with a value of 6.3 ± 2.4, while it is the [l]-type interaction which gives the same effect in the measure *number of ejaculations* (6.0 ± 2.0). Tests of inhomogeneity of the P_1, P_2, and F_1 variances show that three of the female measures, *latency of estrus*, *duration of maximum lordosis* and *frequency of mounting* and two male measures, *circling* and *nuzzling*, exhibit significant genotype-environmental interactions. In all, therefore, the data from six measures would require rescaling to remove either genic or genotype-environmental interaction in order to proceed with the analysis of the second degree statistics.

Unfortunately, the observation on *latency of estrus* in the females are not presented in a manner which allows rescaling. For the other measures both square root and log transformations were made and the latter in all cases removed both sources of non-independence. Thus, for the two measures showing genic interaction the joint scaling tests on the log transformed data gave values of $[d] = 0.138 \pm 0.014$, $[h] = 0.030 \pm 0.028$ for *duration of maximum lordosis*, and $[d] = 0.012 \pm 0.009$, $[h] = 0.126 \pm 0.018$ for *number of ejaculations* both of which now gave satisfactory fits with the observed data. A solution for D, H, and E_1 was therefore attempted from the second degree statistics of all the measures on all scales. These estimates proved to be disappointing. Only two measures, *frequency of mounting* in the females and *intromissions* in the males gave evidence of segregation in the F_2 and backcross generations on any of the three scales employed—in the former case, on two of them. That is, only for these two measures were the magnitudes of V_{F_2}, and $V_{B_1} + V_{B_2}$ greater than the estimates of their environmental component (E_1 and $2E_1$, respectively) ob-

TABLE 5

Jakway's Data: Means and Their Standard Errors in Units as Indicated and (n)

Measure	Generation Means							
						B_1 Reciprocals		
	P_1 (strain 2) (20)	P_2 (strain 13) (18)	F_1 (20)	F_2 (30)	B_1 (53)	P_1 females (27)	F_1 females (26)	B_2 (58)
Circling/15 seconds	0.7 ±0.05	0.3 ±0.03		0.4 ±0.03	0.5 ±0.03	0.4±0.02	0.6±0.03	0.4 ±0.02
Nuzzling/15 seconds	0.7 ±0.04	0.5 ±0.04	0.4±0.02	0.6 ±0.03	0.7 ±0.02	0.6±0.03	0.7±0.03	0.6 ±0.02
Mounting/15 seconds	0.8 ±0.04	0.3 ±0.03	0.5±0.02	0.5 ±0.02	0.6 ±0.03			0.3 ±0.02
Intromissions/15 seconds	0.06±0.01	0.04±0.01	0.1±0.01	0.08±0.09	0.07±0.007			0.06±0.005
Ejaculations in 7 tests	1.1 ±0.3	0.6 ±0.2	4.9±0.4	1.5 ±0.3	1.7 ±0.3	2.5±0.4	0.9±0.20	1.1 ±0.2
Sexual behavior score	6.4 ±0.2	4.6 ±0.2	7.3±0.3	5.7 ±0.2	6.2 ±0.2			4.9 ±0.1

TABLE 6

JAKWAY AND GOY'S DATA: COMPONENTS OF MEANS AND THEIR STANDARD ERRORS FROM JOINT SCALING TESTS, AND HERITABILITIES

Measure	Components of means		Interaction p from χ^2	Heritability [d]/([d]+e) (%)
	[d]	[h]		
Females				
Latency of estrus (hours)	1.2 ±0.1	−0.88±0.21		87.5
Duration of estrus (hours)	1.6 ±0.2	0.99±0.30		87.2
Duration of maximum lordosis (seconds)	7.3 ±0.2	−3.17±1.16	0.001	89.2
Frequency of mounting	8.7 ±0.5	−4.71±0.88		
Males				
Circling/15 seconds	0.18±0.06	−0.17±0.18		83.4
Nuzzling/15 seconds	0.10±0.02	−0.16±0.04		77.6
Mounting/15 seconds	0.2 ±0.02	−0.12±0.03		89.4
Intromissions/15 seconds	0.01±0.005	0.09±0.01		
Ejaculations in 7 tests	0.34±0.18	3.7 ±0.4	0.001	54.3
Sexual behavior score	0.98±0.13	1.6 ±0.2		83.1

tained from the parental and F_1 variances. The results of these estimations are given in Table 7.

Estimates of heritability are therefore only possible for *frequency of mounting* and *intromissions* and these lie between 50% and 60% (see Table 7). An estimate of the number of effective factors (K) is also confined to these two measures neither of which give values greater than one, although it is quite clear that more

TABLE 7

JAKWAY AND GOY'S DATA: COMPONENTS OF VARIATION AND HERITABILITY

Measure	Scale	Components			Heritability	
		D	H	E_1	$\frac{D}{D+E_1}$ (%)	$\frac{\frac{1}{2}D+\frac{1}{4}H}{\frac{1}{2}D+\frac{1}{4}H+E_1}$ (%)
Frequency of mounting (females)	Linear	26.61	13.52	16.68	61.4	50.0
	Log	0.102	0.246	0.088	53.5	56.0
Intromission	Linear	0.0014	0.0016	0.0011	56.0	50.0

than one gene must be involved in the inheritance of the characters which show genic interaction. To obtain some idea of the heritability of the remaining measures a first degree equivalent of one of our estimates has been evaluated, namely, [d]/([d]+e) where e is the mean standard error derived from the sampling errors of the generation means as follows:[9]

$$e=\sqrt{(V_{\bar{P}_1}+V_{\bar{P}_2}+V_{\bar{F}_1}+V_{\bar{F}_2}+V_{\bar{B}_1}+V_{\bar{B}_2})/6}$$

and the results are indicated in Table 6.

The failure of the second degree statistics to show even evidence of segregation for 8 of the 10 measures analyzed requires further comment. A number of contributory factors are present which might reduce or bias our estimates of these statistics as dedrived from our transformations of the data as published. These include: (*a*) grouping of the data into as few as five classes, (*b*) the use of metrics which place over 50% of the individuals into the zero class, and (*c*) grouping all scores higher than an upper limit set by the higher parent or F_1 into one class which may contain 30% of the individuals of a segregating generation.

The exact consequences of these procedures are difficult to ascertain but it is clear that they lead to scalar problems which have not been resolved by either square root or log transformations, and they could easily reduce the variances of the segregating generations (F_2, B_1, and B_2) to those for the nonsegregating generations (P_1, P_2, and F_1).

Our conclusions are therefore necessarily drawn mainly from the analysis of the first degree statistics and are substantially in agreement with those of Goy and Jakway (1959) and Jakway (1959). There are, however, differences in detail. To take one example; they make no allowance for the possibility of genic interaction which is unambiguously present in two of the measures on the original scale. In consequence, they consider that *maximum lordosis* in females is under the control of a single genetic factor without dominance; an interpretation which is difficult to uphold in view of our demonstration of significant [j]-type nonallelic interactions.

Scott

We can now turn to the analysis of less complete sets of data, the most recent among which is that of Scott (1954). This is one small segment of that which has been collected at Bar Harbor on the performance of five thoroughbred strains of dogs. The two strains for which crossbred data have so far been reported are the cocker spaniel and the African basenji or barkless dog. The breeding program used is described as follows:

> It was found that all of the breeds showed a great deal of variability, a large part of which appeared to be hereditary since offspring of different matings gave different results. In order to reduce this variability somewhat, the animals chosen from the parent strains for the crossbreeding experiment were descended from one brother×sister mating in the basenji's and from two matings of a single male with his sister and mother in the case of the cocker spaniels. No selection of these individuals was used except that the original pairs were vigorous and healthy animals. As it turned out later these did not necessarily illustrate the extremes of either breed in all characteristics. Reciprocal crosses were made between these two groups of siblings and an effort was made to obtain at least four dif-

[9] This formula is only applicable in this case because of the demonstrated absence of heritable variation in the segregating F_2 and backcross generations. These may therefore be treated as equivalent to the parental and F_1 generations in displaying only environmentally induced variation, although, of course, they would normally be excluded from this type of estimate.

> ferent pairs in each case, giving two F_1 populations. F_1 males were backcrossed to the mothers so that backcross and F_1 animals raised by the same mothers could be paired. Finally, F_2 populations are being obtained from both crosses (Scott 1954, p. 745).

The subjects were reared in a carefully standardized manner as part of the program, and subjected to a battery of psychological and physiological tests at various predetermined stages in their life history. Scott (1954) gives the results of analyses involving nine different measures, but only in one case is the grouped data for the distributions of the scores given, thus enabling calculations of the approximate means and variances for the various generations. These were the scores derived from a barrier test given the pups at the age of 6 weeks on 2 days. The task is to seek the way round a barrier to reach the experimenter and food.

The F_2 data have not yet been published, so that the results to be found in Table 8 relate to the parental, F_1 and backcross lines only. With both backcross means higher than either of the parental or F_1 means, no scale is possible on which genic interaction is absent. It is not surprising, therefore, to find significant deviations from additivity of gene action on the two available scaling tests, the A and B tests ($p = 0.01$ and 0.05, respectively). Nor is this situation improved by a square root or a log transformation.

In the absence of the F_2 generation mean we can only estimate [d] and [j] among the components of the means. In such a situation, we can, however, obtain estimates of various *compounds* of the remaining components. Thus,

$$[i]-[j]=\bar{F}_1-\tfrac{1}{2}(\bar{P}_1+\bar{P}_2)$$

$$[i]+[l]=\bar{P}_1+\bar{P}_2+2\bar{F}_1-2\bar{B}_1-2\bar{B}_2$$

$$[h]+[l]=\tfrac{1}{2}\bar{P}_1+\tfrac{1}{2}\bar{P}_2+3\bar{F}_1-2\bar{B}_1-2\bar{B}_2$$

Scott's data give the following values for these components on the linear scale:

$$[d]= 1.7\pm 3.9$$

$$[j]=-4.9\pm 8.4$$

$$[h]-[i]=-1.3\pm 1.9$$

$$[i]+[l]=-30.5\pm 8.6$$

$$[h]+[l]=-31.8\pm 8.6$$

Only [i]+[l] and [h]+[l] are significant; therefore the only certain feature of these data is the presence of nonallelic interactions.

From the second degree statistics we can estimate only D+H, $\Sigma(dh)$, and E_1, but we cannot place much reliance on their values in the presence of both unscalable nonallelic interactions and slight genotype-environmental interaction ($p = 0.05$).

TABLE 8

SCOTT'S DATA: MEANS AND THEIR STANDARD ERRORS IN NUMBER OF ERRORS AND (*n*)

	Generation Means				
	P_1 Basenji	P_2 Cocker Spaniel	F_1	B_1	B_2
Actual Parents	2.5 ± 1.7 (16)	10.8 ± 2.3 (26)	5.4 ± 1.0 (41)	12.8 ± 2.7 (27)	14.5 ± 2.9 (23)
Total Population	3.2 ± 1.0 (39)	12.5 ± 1.7 (49)			

A square root transformation removes the latter without, as has been noted, removing the genic interaction. On this scale:

$$D+H=0.329$$

$$\Sigma(dh)=0.008$$

and $$E_1=0.084$$

Although we cannot separate D and H, we can estimate the likely order of heritability by putting D = H and H = 0 in turn which lead to estimates of 66% and 80%, respectively. The value of Σ(dh) provides no evidence of dominance and if the genes show dominance the dominant alleles must be equally distributed between the two parents. With a nonsignificant estimate of [d], on all scales tried, no estimate of the number of effective factors has been attempted. The presence of genic interactions, however, shows that a number of genes are involved.

Tryon

A further set of data are provided by Tryon (1929, 1940, 1942), whose study of selective breeding for "maze-brightness" and "maze-dullness" in rats is probably the best known in the whole of psychogenetics. He selected through 22 generations of brother×sister mating for high and low error scores in a 17-unit automatic maze in which the rats were trained for 19 trials to run to a food reward. He claims that

> Rigorous environmental controls were effected (1) by instituting standard procedure of animal care and of breeding, (2) by using an automatic mechanical device for delivering the animals into the maze without handling, and (3) employing an electric recorder for the scoring of each rat's maze run (Tryon, 1940, p. 112).

Elsewhere Tryon (1931) details the husbandry and comments as follows:

> Very special efforts were made to keep ambient influences the same for all the cages in which these animals lived before they learned the maze (Tryon, 1929). Each animal lived with its siblings until shortly after weaning time (30 days), when it was numbered by punching its ears. Then it was placed with 4 animals from other litters in a cage in which it lived until it ran the maze. Each living cage possessed an ever present supply of food and water. All cages were cleaned at the same time and in the same manner. Even so, it would be naïve to suppose that life withing a cage was *identical* for all animals. Any rat experimenter knows that social life within a cage is variable and complex. But it would not seem likely that the difference in experience of different rats in the same cage would to any significant degree *cause* differences in the later learning of the maze under the remote solitary experimental conditions (p. 316).

A cross was made between the two strains developed, and the F_1 and F_2 generations bred, though it is not clear at what stage in the selection experiment this was done. Tryon (1940) only gives the results in the form of histograms showing the percentage of subjects having a particular error score on his "normalized" scale but from these it has been possible to make approximate reconstructions of the original distributions. The results obtained in this way are given in Table 9.

In the absence of the backcrosses we can apply only the C scaling test

TABLE 9

TRYON'S DATA: MEANS AND THEIR STANDARD ERRORS IN NUMBER OF ERRORS AND (*n*)

Generation Means			
P_1 (bright)	P_2 (dull)	F_1	F_2
25.9±0.9 (85)	142.9±3.7 (53)	63.1±3.5 (133)	71.2±2.9 (202)

and this suggests that on the chosen scale genic interactions are absent. We can therefore estimate [d] and [h] in the manner discussed in the next section and which give values of 117.0 ± 3.8 and −21.3 ± 4.0, respectively. Thus we have a large significant additive effect and a small but significant dominance contribution.

Unfortunately, there is significant genotype-environmental interaction ($p<0.01$), which, in the absence of significance in the C scaling test, we will not attempt to scale out.[10] Our analysis of the second degree statistics is thus prospectively biased. In any case only estimates of $\frac{1}{2}D+\frac{1}{4}H = 646.63$, and $E_1 = 1006.01$ can be obtained, giving the percentage of heritable variation in the F_2 population as 39. If we assume $D=H$ or $H=0$, we can (*a*) estimate the number of effective factors as 14.1 and 10.6, and (*b*) obtain values for our other index of heritability of 49% and 56%, for these two situations, respectively.

Thus the pattern of rat behavior investigated by Tryon is controlled by many genes which are additive in their effect but interact with the environment. They show dominance and there is a preponderance of dominant genes for a low score. The heritabilities for this character are about average.

Vicari

A further set of data which omit the backcrosses are provided by the investigations of Vicari (1929). These are the earliest psychogenetical experiments which yield any data amenable to analysis by the methods proposed here. She used four strains of mice which were "closely inbred" and derived F_1 and F_2 generations from them. One of these strains, the Japanese waltzer, was regarded at the time as being of a different species from the other three (*Mus musculus*), and the offspring of the cross involving it consequently hybrids. It has since been shown (see Grüneberg, 1952), however, that the Japanese waltzer is a subspecies of *Mus musculus*, and furthermore, that it differs from the normal in a genetically complex manner, that is, the waltzer condition is not due to a single gene difference. Vicari's measures were derived from a simple, two-choice maze in which the subjects ran to a food reward. No details of the deprivation schedule for motivating the animals to run this maze are given, and it is clear from the running times reported that the apparatus itself was ill-designed for the purpose of obtaining efficient learning. Despite these difficulties, however, Vicari reports the results after 14 trials for a substantial number of subjects from the four parental generations and the three F_1s and F_2s bred from them. We can, of course, only apply the C scaling test, the results of which are given along with the generation means in Table 10. Where there is a significant deviation on the scaling test only estimates of the following *compounds* and their standard errors can be obtained,

$$[d]-\tfrac{1}{2}[j]=\tfrac{1}{2}(\overline{P}_1-\overline{P}_2)$$

and

$$[h]-[i]=\overline{F}_1-\tfrac{1}{2}(\overline{P}_1+\overline{P}_2)$$

as before, and from Scaling Test C itself:

$$2[i]+[l]=\overline{P}_1+\overline{P}_2+2\overline{F}_1-4\overline{F}_2$$

[10] Since, for most of the data previously considered in this paper, rescaling has been necessary because of a failure of *both* scaling criteria, no attempt has been made to scale out the significant genotype-environmental interaction in this case on the grounds that improvement in respect of the latter disturbance might be at the expense of the satisfactory outcome of the test for additivity.

TABLE 10

VICARI'S DATA: MEANS AND THEIR STANDARD ERRORS OF MEASURES SHOWN AND (n)

Measure	Quantity	Generation Means									
		Japanese Waltzer	Cross $P_1 \times P_2$		Albino	Cross $P_2 \times P_3$		Dilute Brown	Cross $P_3 \times P_4$		Brown (Abnormal eyed) P_4
		P_1 (28)	F_1 (119)	F_2 (49)	P_2 (78)	F_1 (31)	F_2 (45)	P_3 (27)	F_1 (14)	F_2 (62)	(26)
Number of trials to first errorless run	Mean	8.8±0.8	9.2±0.4	10.5±0.5	8.7±0.4	7.2±0.7	5.5±0.9	6.9±0.7	6.1±0.6	8.1±0.5	10.1±0.8
	C Scaling Test[a]		−6.1 ±1.6**			8.1 ±2.5**			−3.3 ±1.4**		
	Genotype-environmental interaction		2.0*			3.6**			1.7*		
	$[d]-\frac{1}{2}[j]$		0.05±0.45			0.9±0.4			1.6 ±0.5		
	$[h]-[i]$		0.45±0.6			−0.6±0.8			−2.4 ±0.8		
Largest number of consecutive errorless runs	Mean	0.9±0.2	2.0±0.2	0.9±0.1	1.4±1.3	1.4±0.2	1.8±0.1	1.5±0.1	1.8±0.2	1.9±0.2	1.1±0.2
	C Scaling Test		2.5±1.4			−1.3±1.4			−1.5±1.3		
	Genotype-environmental interaction		117.7**			244.1**			2.4		
	$[d]-\frac{1}{2}[j]$		0.25±0.7			0.05±0.65			0.2±0.1		
	$[h]-[i]$		0.85±0.3			−0.05±0.7			0.5±0.2		
Total number of errorless runs	Mean	1.5±0.3	2.7±0.3	1.5±0.3	1.9±0.2	2.6±0.4	3.1±0.3	2.8±0.3	3.2±0.6	2.7±0.3	1.7±0.4
	C Scaling Test		2.9±0.8**			−2.5±0.5**			0±1.5		
	Genotype-environmental interaction		3.7**			1.7			2.1		
	$[d]-\frac{1}{2}[j]$		0.2±0.15			0.45±0.2			0.55±0.25		
	$[h]-[i]$		1.0±0.4			0.25±0.4			0.95±0.65		
Mean time per run in seconds	Mean	73.1±8.6	41.9±3.7	70.7±8.5	50.1±4.4	15.8±1.6	19.1±2.2	12.5±1.7	13.9±2.1	12.9±2.1	27.0±2.9
	C Scaling Test		−75.9±20.9**			16.9±7.2*			15.5±5.8*		
	Genotype-environmental interaction		1.4			19.2**			3.5**		
	$[d]-\frac{1}{2}[j]$		11.5±4.8			18.8±2.4			7.3±1.7		
	$[h]-[i]$		−19.7±5.9			−15.5±2.9			−5.9±2.7		

[a] 2[i]+[1] (see text).

* Significant at the 5% level.

** Significant at the 1% level.

TABLE 11

VICARI'S DATA FOR JAPANESE WALTZER × ALBINO CROSS: MEANS AND THEIR STANDARD ERRORS IN MEAN RUNNING TIME AND (n)

Trial	Generation Means				C Scaling Test	Genotype-environmental interaction
	Japanese Waltzer	Cross $P_1 \times P_2$		Albino		
	P_1	F_1	F_2	P_2		
1	113.7±12.0 (80)	74.2±7.4 (151)	86.8±11.7 (61)	92.3±11.0 (80)	7.2±32.2	1.5**
4	73.4± 8.6 (68)	67.3±7.2 (149)	70.6±11.4 (61)	60.2± 8.7 (79)	−12.0±29.7	1.5**
8	52.3±10.0 (45)	47.0±5.9 (128)	60.3±10.0 (53)	47.1± 5.7 (78)	−47.7±26.0	1.3*
14	83.0±16.8 (28)	21.1±2.6 (119)	51.7±11.6 (49)	29.6± 4.6 (78)	−52.2±29.4	9.7**

* Significant at the 5% level.
** Significant beyond the 1% level.

But direct estimates of [d] and [h] can be obtained where there is no significant deviation on the scaling test by assuming [i] and [j] = 0, as in the case of Tryon's data above.

Of the 12 sets of data, i.e., four measures recorded for each of the three crosses, eight show significant genic-interaction and eight, genotype-environmental interaction on the chosen scales. Unfortunately we cannot rescale the data because Vicari does not give the individual scores on which the means and variances are based. She does, however, present the distributions for one measure, mean running time, for different stages in the experiment, the first, fourth, eighth and fourteenth trials. We shall use the cross $P_1 \times P_2$ to illustrate the further analyses, the relevant generation means appearing in Table 11. The decrease in the number of subjects in successive trials which is observed is attributed by Vicari to death, escapes, failures to run, etc. An analysis of variance of the 4×4 table, that is, four generation means in each of four trials, shows a highly significant difference between trials ($p < 0.01$) and a significant difference between generations ($p = 0.05$–0.01) when compared with the interaction mean square for generations × trials, which has the same order of magnitude as the sampling variance of the generation means. The significant difference between generations is expected if the character is inherited and the significant difference between trials, the running time falling steadily as the number of trials increases, reflects a strong training component which is presumably non-heritable. Since the C scaling test when applied to these data detected no significant deviations due to genic-interaction (Table 11), we can estimate [d] and [h] directly from the generation means on the original scale.[11] For the first trial $[d] = 10.7 \pm 8.2$ and $[h] = -28.8 \pm 17.8$ neither of which are significant, that is, there is no significant heritable component of the generation means. The fully trained performance in the fourteenth trial gives $[d] = 26.7 \pm 8.7$ and $[h] = -35.2 \pm 9.1$. The components are similar in absolute and relative magnitudes in these two extreme cases but their significance is higher in the last trial This greater heritability in the last trial is supported by the second degree statistics. Thus in the first trial the percentage of heritable variation in the F_2 population

[11] In ignoring the significant genotype-environmental interactions shown in Table 11, we are following the argument of Footnote 9.

is not significantly different from zero, while in the fourteenth trial it is 51%. Furthermore the fourth and eighth trials in this case give intermediate values of 21% and 29%, respectively. Hence heritability increases almost linearly with the number of trials, the final performance being more heritable than the initial performance.

Before attempting an interpretation of this finding two points must be borne in mind. Firstly, the experimenter's skill may have improved in successive trials, thus reducing the nonheritable component of variation, though this is unlikely since the generations were probably not all tested at the same time, and, secondly, almost 30% of the animals scored in the first trial were missing in the fourteenth for a number of reasons. If the missing third were not a random sample of the original subjects a progressive bias could have been introduced. It is not possible to ascertain from the data available whether or not one or both of these factors is making a contribution to the observed trend. It is interesting, however, to note that the total variation in the F_2 population remained constant from the first to the last trial and the increased heritability results from a drop in the percentage due to nonheritable agencies from 100 to 49. It seems likely, therefore, that what we have detected is in fact a real effect and that it represents a progressive release of the performance from the effect of environmental stimuli irrelevant to it.

This effect of a progressive increase in the heritability of performance indicated in Vicari's data would seem, if confirmed, to have interesting implications. It might be related to the change over from general to specific factors known to occur in the acquisition of skills (Fleishman, 1957; see also Wherry, 1939), and might also indicate a method for assessing the relative importance of environmental variation in learning tasks. The effect of the *same* environmental stimuli at different stages in a given task might be studied, as in Vicari's situation, or at the same stage in different tasks, as well as that of *different* stimuli in either of such arrangements. As was noted earlier, the sort of problems which the inclusion of environmental variation introduces into biometrical analyses has been discussed and possible solutions indicated (Jones & Mather, 1958; Mather & Jones, 1958; van der Veen, 1959). While it is beyond the scope of this paper to enter into this matter in detail, it may be said that there is evidence for a genetical component in the determination of the *variability* of performance in such different environments, as opposed to the control of its actual expression in any single one of them which is what we have been dealing with so far. This variability is also susceptible to analysis by biometrical methods (Jinks & Mather, 1955). The analysis from this point of view of the only suitable behavioral data at present known to us (Broadhurst, 1960) is not yet complete.

Thus, while the nature of Vicari's data makes our conclusions tentative, the analyses discussed here have shown their advantages if only in indicating the complexity of the inheritance of the behavior patterns under investigation.

Discussion

We have presented the results of our reanalyses of all the data available to us in the field of psychogenetics with little reference to the outcome of analyses of other kinds performed, in some cases, by the authors concerned. The methods used have usually been those of classical Mendelian genetics, which, though basic to biometrical genetics, cannot satisfactorily be applied in their simpler forms to the analysis of continuously variable characteristics. The clarity associated with the Mendelian analysis of discontinuous variation and attributable thereby to the effects of two or three major genes is not to be expected from biometrical methods, a major assumption of which is that the continuous phenotypic variation observed is the product of multiple genetical and environmental causes, largely unspecifiable in detail. Sometimes the argument from Mendelian analysis has been by analogy, which is not always illuminating and may in certain instances be downright misleading. For example, the resemblance of the F_1 to one or other of the parental lines does not necessarily mean the same thing in biometrical genetics as it does in the simpler cases encountered in Mendelian analysis. According to the polygenic hypothesis, the departure of the F_1 mean from the mid-parental value depends both on the balance in the parental lines of the dominant and recessive polygenes and upon the respective *direction* of their cumulative effects. Dominants may be increasers or decreasers, that is, having a positive or a negative phenotypic effect, respectively, as expressed on the scale used, or there may be a balance between the two. Thus the F_1 mean value may be close to the upper parental mean on a scale because of a preponderance in the parents of dominants with positive effect in terms of the metric, while an F_1 close to the lower parental value will result from a preponderance of dominant decreasers. Intermediate values may express different degrees of balance operating. Failure to recognize this difference between potence, measured in terms of [h], and dominance, measured as $\sqrt{H}$, will in general lead to spurious disagreements between the level of dominance in the F_1s and that in the segregating F_2 and backcross generations (e.g., Jakway, 1959, p. 155). It is equally misleading to regard low potence, that is an F_1 close to the mean parental value, as a diagnostic characteristic of multifactorial inheritance (see Hall, 1951, p. 321). Genetical methods and analyses more complex than those that have been considered here, such as the analysis of double back-crossing (Mather, 1949), or of diallel crosses (Broadhurst, 1959, 1960; Hayman, 1954; Jinks, 1954), are needed to enable more precise estimates to be made of the different factors operating. In this connexion it should be noted that the analysis in turn of single crosses from two parental strains such as we have been dealing with here is a relatively inefficient and laborious method of making a biometrical analysis of quantitative data. Techniques involving crossing several pure strains at once are superior, and of these the diallel cross method in which a number of strains are intercrossed reciprocally in all possible combinations and analyzed, together with the parental lines, in a single diallel table is probably the best. The merits of this approach at the present time in psychogenetics have been argued elsewhere (Broadhurst, 1960); they are, briefly, that the analysis can proceed at the F_1 stage, without the necessity of breeding further generations, and that the method thereby provides a quick survey in several parental strains at once of the genetical determinants of the character investigated. The former may be of particular utility in interspecies crosses, where hybrids are sometimes sterile, so precluding the possibility of breeding F_2s, etc. Intensive study of the gene differences in pairs of strains selected for further analysis in this way can then follow.

One source of satisfaction from the present work has been the consistently high heritabilities obtained from our analyses; and by choosing a scale, wherever possible, on which interactions between genes, and between genes and environment are absent, these estimates are somewhat more reliable as well as higher than those obtained by the authors of the papers under review. Any inefficiencies in the experimental design—e.g., the imperfect control of environmental variation, or the unreliability of the measures used—or in the statistical analysis used will reduce the values obtained below their true values. While it has been beyond our power to influence the first source of inefficiency in these data, our analyses have improved the position regarding the second. And the collaboration between psychologist and geneticist will ensure that future researches are designed with due care and attention to the importance of the relevant genetical principles (see Broadhurst, 1960).

The choice of scale is an important problem in biometrical analysis (see Mather, 1949) and, as we have seen, the need for rescaling, resulting from interactions between the genes and between the genes and the environment, arises in the inheritance of some 50% and 70%, respectively, of the measures in the examples reviewed here. In no case was the presence of genic interaction demonstrated as a significant factor prior to the analyses undertaken for this review and in only one case were steps taken to eliminate the gene-environment type of interaction by a scalar change (Thompson & Fuller, see Footnote 3). And yet our analyses show that in almost all cases a simple log transformation is sufficient to eliminate both causes of interaction and hence provide a scale on which unambiguous interpretations of dominance and potence effects can be made.

The gene-environment interaction detected by the inhomogeneity of the variances of the parents and F_1s has two main causes. The major cause (some 60% of measures) is a correlation between mean and variance in the nonsegregating generations. A more interesting cause, however, in the remaining examples is the lower variance of the F_1 individuals compared with those of the parental generations; an effect which is independent of their means. This phenomenon, which is common to the inheritance of all types of characters and occurs equally among animals and plants, has received considerable attention of late (see Jinks & Mather, 1955; Lerner, 1954; Mather, 1953, for reviews). Extensive discussion of this point is, however, beyond the scope of the present review.

Some 70% of the genic interactions in these analyses are due to the [j]-type interactions. This has two implications. Firstly, there must be interactions between additive and dominance effects and, secondly, the interacting genes must be associated in the parental lines, the majority of increasing interacting genes being present in one parent and the decreasers in the other. This is not an unexpected result when one considers that most of the experiments reviewed here have employed parental lines which were chosen because they represented the extreme phenotypes immediately available or obtainable as a result of prolonged selection.

This policy could explain two further features of these examples, namely, the rarity of heterosis and the often satisfactory estimates of the number of genetical factors attained by a method which, for reasons mentioned earlier and discussed more fully by Mather (1949), have often failed to give sensible values in other work. Without going into details it is clear that if the better parent in a cross already contains the majority of the available increasing genes it is unlikely to give rise to a superior F_1 irrespective of the dominance or interactive properties of the genes. Similarly, our estimate of the number of genetical factors assumes that the genes are associated in the parental lines. Failure of this assumption leads to underestimation. Our rather satisfactory estimates could, therefore, be a further indication that the genes are so distributed in the parental lines.

It is hoped that the reanalyses reported here serve as another example of the application of biometrical methods to psychological data in addition to those already available (Broadhurst, 1959, 1960). Considering the unsatisfactory nature of much of the data at our disposal, it is felt that the outcome, in terms of the ease of the analyses, especially with regard to the search for suitable scales, and the consistency of the results obtained, has been favorable. It is not yet possible to pronounce on the general efficiency in this field of the methods advocated by us. Further proof of their suitability will only come when they have been applied more widely to data gathered from suitably designed experiments, perhaps along the lines indicated, so that replication of the genetical picture becomes possible, thus enabling some specification to be made of the generality of the determinants of a particular behavioral characteristic.

Summary

The techniques which can be used in the analysis of quantitative data by the methods of biometrical genetics were outlined and the importance of achieving a suitable scale noted. The body of the paper consists of descriptions of experiments in psychogenetics which lend themselves to this type of analysis, and a presentation of the results of our reanalyses of the data they provide in terms of additive, dominance and interaction components of variation. We conclude that, despite the unsuitable nature of some of the available data, the outcome indicates the utility of the biometrical approach.

REFERENCES

Broadhurst, P. L. Application of biometrical genetics to behaviour in rats. *Nature, Lond.*, 1959, **184**, 1517–1518.

Broadhurst, P. L. Experiments in psychogenetics: Applications of biometrical genetics to the inheritance of behaviour. In H. J. Eysenck, (Ed.), *Experiments in personality:* Vol. I. *Psychogenetics and psychopharmacology.* London: Routledge & Kegan Paul, 1960. Pp. 1–102.

Broadhurst, P. L. Analysis of maternal effects in the inheritance of behavior. *Anim. Behav.*, 1961, in press.

Brody, Elisabeth G. Genetic basis of spontaneous activity in the albino rat. *Comp. Psychol. Monogr.*, 1942, **17**, No. 5, 1–24.

Cavalli, L. L. An analysis of linkage in quantitative inheritance. In E. C. R. Reeve & C. H. Waddington (Eds.), *Quantitative inheritance.* London: Her Majesty's Stationery Office, 1952, Pp. 135–144.

Dawson, W. M. Inheritance of wildness and tameness in mice. *Genetics*, 1932, **17**, 296–326.

Fleishman, E. A. A comparative study of aptitude patterns in unskilled and skilled psychomotor performances. *J. appl. Psychol.*, 1957, **41**, 263–272.

Fuller, J. L., & Thompson, W. R. *Behavior genetics.* New York: Wiley, 1960.

Goy, R. W., & Jakway, Jacqueline S. The inheritance of patterns of sexual behaviour in female guinea pigs. *Anim. Behav.*, 1959, **7**, 142–149.

Goy, R. W., & Young, W. C. Strain differences in the behavioral responses of female guinea pigs to alpha-estradiol benzoate and progesterone. *Behaviour*, 1957, **10**, 340–354.

Grüneberg, H. *The genetics of the mouse.* (2nd ed.) The Hague: Martinus Nijhoff, 1952.

Hall, C. S. The genetics of behavior. In S. S. Stevens (Ed.), *Handbook cf experipxperimental psychology.* New York: Wiley, 1951. Pp. 304–329.

Hayman, B. I. The theory and analysis of of diallel crosses. *Genetics*, 1954, **39**, 789–809.

Hayman, B. I., & Mather, K. The description of genic interactions in continuous variation. *Biometrics*, 1955, **11**, 69–82.

Jakway, Jacqueline S. The inheritance of patterns of mating behaviour in the male guinea pig. *Anim. Behav.*, 1959, **7**, 150–162.

Jinks, J. L. The analysis of continuous variation in a diallel cross of *Nicctiana rustica* varieties: I. The analysis of F_1 data. *Genetics*, 1954, **39**, 767–788.

Jinks, J. L., & Jones, R. M. Estimation of the components of heterosis. *Genetics*, 1958, **43**, 223–234.

Jinks, J. L., & Mather, K. Stability in development of heterozygotes and homozygotes. *Proc. Roy. Soc. Lond., Ser. B.*, 1955, **143**, 561–578.

Jones, R. M., & Mather, K. Interaction of genotype and environment in continuous variation: II. Analysis. *Biometrics*, 1958, **14**, 489–498.

Lerner, I. M. *Genetic homeostasis.* Edinburgh: Oliver & Boyd, 1954.

Loeb, L., & Wright, S. Transplantation and individuality differentials in inbred families of guinea pigs. *Amer. J. Pathol.*, 1927, **3**, 251–283.

Mather, K. *Biometrical genetics: The study of continuous variation.* London: Methuen, 1949.

Mather, K. Genetical control of stability in development. *Heredity*, 1953, **7**, 297–336.

Mather, K., & Jones, R. M. Interaction of genotype and environment in continuous variation: I. Descriptions. *Biometrics*, 1958, **14**, 343–359.

Pearson, E. S., & Hartley, H. O. *Biometrika tables for statisticians.* Vol. I. (2nd ed.) Cambridge: Univer. Press, 1958.

Rundquist, E. A. Inheritance of spontaneous activity in rats. *J. comp. Psychol.*, 1933, **16**, 415–438.

Scott, J. P. The effects of selection and domestication upon the behavior of the dog. *J. Nat. Cancer Res. Inst.*, 1954, **15**, 739–758.

Snedecor, G. W. *Statistical methods applied to experiments in agriculture and biology.* (5th ed.) Ames: Iowa State Coll. Press, 1956.

Thompson, W. R. The inheritance of behaviour: Behavioural differences in fifteen mouse stains. *Canad. J. Psychol.*, 1953, **7**, 145–155.

Thompson, W. R. The inheritance of behavior: Activity differences in five inbred mouse strains. *J. Hered.*, 1956, **47**, 147–148.

Tryon, R. C. Genetics of learning ability in rats: Preliminary report. *U. Calif. Publ. Psychol.*, 1929, **4**, 71–89.

Tryon, R. C. Studies in individual differences in maze ability: IV. The constancy of individual differences: Correlation between learning and relearning. *J. comp. Psychol.*, 1931, **12**, 303–345.

Tryon, R. C. Genetic differences in maze-learning ability in rats. *Yearb. Nat. Soc. Stud. Educ.*, 1940, **39**(1), 111–119.

Tryon, R. C. Individual differences. In F. A. Moss (Ed.), *Comparative psychology.* (2nd ed.) New York: Prentice-Hall, 1942. Pp. 330–365.

van der Veen, J. H. The 2 × 2 genotype-environment table. *Heredity*, 1959, **13**, 123–126.

Vicari, Emelia M. Mode of inheritance of reaction time and degree of learning in mice. *J. exp. Zool.*, 1929, **54**, 31–88.

Wherry, R. J. Factorial analysis of learning dynamics in animals. *J. comp. Psychol.*, 1939, **28**, 263–272.

Wigan, L. G. Balance and potence in natural populations. *J. Genet.*, 1944, **46**, 150–160.

Wright, S. The genetics of quantitative variability. In E. C. R. Reeve & C. H. Waddington (Eds.), *Quantitative inheritance.* London: Her Majesty's Stationery Office, 1952. Pp. 5–41.

(Received June 9, 1960)

Psychological Bulletin
1961, Vol. 58, No. 5, 363-375

TEACHING MACHINES: A REVIEW[1]

CHARLES S. MORRILL[2]
MITRE Corporation

The same forces which have characterized the evolution of general educational practices are inherent to the history of the new science of automated teaching. As a result of the expansion and multiplying complexities of political, economic, and social interests, there developed an ever increasing need for the rapid education of large numbers of people. New educational objectives demanded new methods of instruction, and the history of education is marked by many diverse attempts at establishing more efficient teaching procedures. Once again teaching methods must be reevaluated. Rigid adherence to the principle of personal teacher-student relationships no longer seems feasible—an instructional system more appropriate for present-day needs must be established. It is probable that the use of automated teaching devices can fill this need in the method of education. As Corrigan (1959) has suggested:

> the automated teaching method has grown out of a pressing need. This need has been created by a twofold technical training problem. As advances in science and technology have been made, there has been an ever increasing demand for well-trained instructors; at the same time the availability of these trained persons has been diminishing. This situation is aggravated further by the increased scope and complexity of subjects, and the ever increasing ratio between number of instructors and students (p. 24).

CURRENT TRENDS IN AUTOMATED TEACHING MACHINES

Current interest in the area of automated teaching machines is well illustrated by the simple index of frequency-per-year of published teaching machine articles. Fry, Bryan, and Rigney (1960) report that for the years prior to 1948 there are only 6 references, whereas through 1959 there were more than 50 reports published.

The grandfather of automated teaching machines is Sydney L. Pressey (1926, 1927), who designed machines for automated teaching during the mid-1920s. His first device was exhibited and described at the American Psychological Association (APA) meetings in 1924; an improved device was exhibited in 1925 at the APA meetings. Both forms of the apparatus automatically performed simultaneous administration and scoring of a test and taught informational and drill material. Pressey's device, about the size of a portable typewriter, presented material to the subject via a small window. Four keys were located alongside the apparatus. If the student activated a key corresponding to the correct answer, the machine advanced to the next item. If his response was incorrect, the machine scored an error and did not advance to the next item until the correct answer was chosen. The capacity of the drum was 30 two-line typewritten items; the paper on which the questions appeared was carried as in a typewriter.

[1] The research reported in this document was supported by the Department of the Air Force under Air Force Contract AF-33-(600)39852.

[2] The author wishes to acknowledge the valuable editorial assistance of Sylvia Pilsucki.

In 1927, Pressey summarized his efforts as follows:

> The paper reports an effort to develop an apparatus for teaching drill material which (a) should keep each question or problem before the learner until he finds the correct answer, (b) should inform him at once regarding the correctness of each response he makes, (c) should continue to put the subject through the series of questions until the entire lesson has been learned, but (d) should eliminate each question from consideration as the correct answer for it has been mastered (p. 552).

In 1930, Peterson devised a self-scoring, immediate feedback device. The Chemo Card, as this device was later called, utilized the technique of multiple choice. A special ink was used by the student in marking his answer. The mark appeared red if the answer was incorrect; a dark color resulted if the answer was correct. Although Pressey's notions and the Chemo Card might have stimulated an interest in automated teaching techniques in the twenties, educators and researchers obviously were not at that time ready for this advanced concept of teaching. Automated teaching did not take hold.

In 1932, Pressey published an article describing a kind of answer sheet which could be scored by an automatic scoring device. This apparatus recorded errors by item, and thus provided the instructor with clues as to what questions needed further instruction. In 1934, Little experimented with this device as well as with the device originated by Pressey in 1926. His results favored the use of automated devices in contrast to regular classroom techniques.

The next appearance of automated teaching literature came a considerable number of years later. During World War II, the Automatic Rater was used by the Navy for training. This device projected a question on a small screen; the subject's response consisted of pushing one of five buttons.

In 1950, Pressey described a new automated device called the Punchboard. Multiple-choice questions were presented to the student. The key answer sheet inside the Punchboard contained holes opposite the correct answers only. If the answer was correct, the student's pencil penetrated deeply; if incorrect, the pencil did not penetrate the paper significantly. Angell and Troyer in 1948 and Angell in 1949 reported the results of using the Punchboard. Both studies suggested the superiority of this method over traditional classroom procedures.

In 1954, Skinner published "The Science of Learning and the Art of Teaching," which provided the basis for the development of his teaching machines. In this article, he stressed the importance of reinforcement in teaching and suggested teaching machines as a method of providing this needed reinforcement for the learner.

Reports concerning the Subject-Matter Trainer began to appear in 1955 (Besnard, Briggs, Mursch, & Walker, 1955; Besnard, Briggs, & Walker, 1955). This electromechanical device is a large multiple-choice machine used essentially for training and testing in the identification of components and in general verbal subject matter. Extensive research has been done with this device because of its considerable flexibility, i.e., it allows several modes of operation for self-instruction: variety of programed subject matter, drop-out feature after items have been mastered, etc.

The Pull-Tab, used experimentally by Bryan and Rigney in 1956, was a device in which the subject received not only a "right" or "wrong" indication after his choice but also a somewhat detailed explanation of "why" a response was incorrect. In 1949, Briggs had found in experimenting with the Punchboard that learning is significantly enhanced by immediate knowledge of results. Bryan and Rigney's data illustrated that the combination of immediate knowledge of results plus explanation, if the student is in error, produced significantly higher scores on a criterion test than if no explanation had been given. The importance of this research from a historical point of view is that it investigated immediate knowledge of results as a factor existing on a continuum with varying degrees of effect. Up to this point any comparison involving the effectiveness of teaching machines had been one between classroom instruction and the "new" machine under consideration. In Briggs' and in Bryan and Rigney's research, however, we see the beginning of a concern, to become greater in the next few years, with the possible effects of specific variables and their interactions on learning.

The years 1957–58 mark the beginning of the period in which resurgent interest in teaching machines was initiated. Ramo's arguments (1957) reopened the consideration of automated techniques for classroom use. His article served as one of the more forceful attempts to alert educators to the needs and requirements for automated techniques in education. Skinner's continued interest (1958) served as the major catalyst in this area. In his article, he reviewed earlier attempts to stimulate interest in teaching machines and further explained that the learning process was now better understood and that this increased sophistication would be reflected in teaching machine technology. Skinner suggested that the most appropriate teaching machine would be that which permits the student to *compose* his response rather than to select it from a set of alternatives. On the basis of this philosophy and in conjunction with other principles of learning theory to which Skinner adheres, he designed a teaching machine with the following characteristics. The questions, printed on a disk, are presented to the student through a window. The student's response is written on a paper tape, which is advanced under a transparent cover when the student lifts a lever. At this point the correct answer appears in the window. If the student is correct, he activates the lever in one manner, which eliminates the item from the next sequence. If he is incorrect, the lever is activated in a different manner, thus retaining the item in the next sequence.

Holland (1960), a co-worker of Skinner's, has suggested several well-known learning principles that should be applied to teaching machine technology: immediate reinforcement for correct answers is a must, learned behavior is possible only when it is *emitted* and reinforced, gradual progression (i.e., small steps in learning sequences and reducing wrong answers) is necessary to establish complex repertoires, gradual withdrawal (fading or vanishing) of stimulus support is effective, it is necessary to control the student's observing and echoic behavior and to train for discrimination, the student should write his response. The Skinner machine does in fact employ these principles.

Ferster and Sapon (1958) described the Cardboard Mask, a most simple teaching machine which employs the principles which Skinner and Holland outline so clearly. This device is a cardboard folder containing mimeographed material which is presented one line at a time. The student, after writing his response on a separate sheet of paper, advances the paper in the mask, thereby exposing the correct response.

In 1958, a number of investigators interested in teaching machines recommended that the programed material be a function of the student's response. This idea suggests that a "wrong" response may not necessarily be negative reinforcement and that both the "right" and "wrong" responses should modify the program. Rath and Anderson (1958) and Rath, Anderson, and Brainerd (1959) have suggested the use of a digital computer which automatically adjusts problem difficulty as a function of the response. Crowder's (1958, 1959a, 1959b) concept of "intrinsic programming" permits the response to alter the programing sequence.

During the last few years, researchers have been focusing their attention on investigating many of the variables which are pertinent to the design and use of teaching machines. The seemingly simple task of defining a teaching machine has been a serious problem to many authors (Day, 1959; Silberman, 1959; Weimer, 1958). Some definitions have made more extensive demands on teaching devices than others. Learning theorists (Kendler, 1959; Porter, 1958; Skinner, 1957; Spence, 1959; Zeaman, 1959) are now most outspoken concerning the application of theoretical concepts to teaching machine technology. Transfer of training, mediational processes, reinforcement, motivation, conditioning, symbolic processes, and language structure are but a few of these areas of interest.

There are indeed many other variables about which there is a divergence of opinion and about which experimental evidence is completely lacking or controversial. The reports of Skinner (1958), Israel (1958), Coulson and Silberman (1960), Fry (1959), and Stephens (1953) are all focused, at least in part, on questions related to response modes, e.g., multiple choice, construction of the response, responses with reinforcement, etc. Briggs, Plashinski, and Jones (1955) investigated self-paced vs. automatically paced machines. The importance of motivation in connection with teaching machines has been explored by Holland (unpublished), Mayer and Westfield (1958), and Mager (1959).

Essentially, the history of automated teaching is short—it started in the mid-twenties and was strenuously reactivated by the appearance of Skinner's 1958 article. Empirical investigations of many important issues in this field are just now beginning to appear. However, the necessity of developing automated teaching methods has been evident for many years.

General Problem Areas

Definition

As in any new field, the first problem is one of definition. What is a teaching machine? Silberman (1959) says that a teaching device consists of four units: an input unit, an output unit, a storage unit, and a control unit. As such, this definition includes a broad category of devices, from the most simple to the most complex. Weimer (1958) goes beyond the device itself, stating that a teaching machine must present information to the student as well as test the student by means of a controlled feedback loop. Crowder (1960) insists that a teaching machine

> must in some way incorporate two-way communication. That is, the student must respond to the information presented by the machine, and the machine must in turn recognize the nature of the student's response and behave appropriately (p. 12).

Perhaps the most inclusive definition is one given by Day (1959):

> A teaching machine is a mechanical device designed to present a particular body of information to the student. . . . Teaching machines differ from all other teaching devices and aids in that they require the active participation of the learner at every step (p. 591).

Although the emphasis in some of the above concepts is different, together they give a rather complete description and, if you will, definition.

Programing

The programing of subject matter for teaching machines is the most extensive and difficult problem in this new technology. Beck (1959) describes specific concepts which he thinks appropriate for programing a Skinner-type machine:

> A student's responses may be restricted and guided in a great number of ways. These range from all types of hints . . . to simply presenting the response which it is desired a student acquire (p. 55).

Carr (1959) discusses in some detail the importance of programing in terms of learning efficiency and retention. Much of what he says remains open for empirical verification. Rothkopf (1960) has suggested that the development of programed instruction suffers from two difficulties: a weak rational basis for program writing and inadequate subject-matter knowledge among program writers.

The extent to which any initial program needs revision is perhaps exemplified by the program in Harvard's course Natural Sciences 114. Holland points out that the first program of materials included 48 disks, each containing 29 frames, whereas a revision and extension of the program the following year included 60 disks of 29 frames each. Holland's objective was to extend the program and decrease the number of student errors. Crowder's (1960) programing objectives are different from Holland's. He states:

> By means of "intrinsic programming" it [the program] recognizes student errors as they occur and corrects them before they can impede understanding of subsequent material or adversely affect motivation (p. 12).

Crowder considers it almost impossible to write a program which completely avoids error, and therefore he structures the program requirements on the probability of error. When an error is made, the next presentation explains the subject's mistake. Depending on the nature of the error and when it occurs, the subject may either return to the original question or enter a program of correctional material.

Another concept for programing is known as *branching* (Bryan & Rigney, 1959). Through branching, many possible routes are provided through which the subject can proceed, depending on the response. The subjects are allowed to skip certain material if they have demonstrated a knowledge of it. One study (Coulson & Silberman, 1960) suggests that under branching conditions subjects require less training time than under nonbranching conditions; however, results on the criterion test were not significantly different.

For certain kinds of subject matter, *vanishing* is still another concept for programing (Skinner, 1958). A complete or nearly complete stimulus is presented to the subject. Subsequent frames gradually omit part of the stimulus until all of it is removed. The subject is then required to reconstruct the stimulus.

To program verbal learning sequences, Homme and Glaser (1959) suggest the Ruleg. With this method, the written program states a rule and provides examples for this rule. In each case, either the rule or the example is incomplete, requiring the subject to complete it.

In a recent study Silverman (1960b) investigated methods of presenting verbal material for use in teaching machines. He recommended that further research involving the design and use of teaching machines should take into consideration the possible use of context cues as a means of facilitating serial rote learning. At the same time, however, he stated that continuous use of context cues as ancillary prompts should be avoided, since such prompts can interfere with learning.

The optimum size of steps and the organization of the programed material are two formidable problems. Skinner (1958) states:

> Each step must be so small that it can always be taken, yet in taking it the student moves somewhat closer to fully competent behavior (p. 2).

In order to determine the value of steps in a program, Gavurin and Donahue (1961) investigated the effects of the organization of the programed material on retention and rate of learning. They state that the assumption that optimum teaching machine programs are those in which items are presented in a logical sequence has been validated for acquisition but not retention. The results of a study carried out by Coulson and Silberman (1959) indicated that small steps were more time consuming but resulted in statistically significant higher test scores on one of the criterion tests. Pressey (1959) in principle disagrees with Skinner's notions of short and easy steps, and he strongly suggests an experimental investigation of this question. Both rate of learning and retention (recall or recognition) are of critical concern.

The above discussion suggests several areas which are directly applicable to programing and which are under investigation and/or need further experimentation. Indeed, there are a number of unanswered questions in the programing complex, some of which have been suggested by Galanter (1959):

1. What is the correct order of presentation of material?
2. Is there an optimum number of errors that should be made?
3. How far apart (in some sense) should adjacent items be spaced?
4. Is experimentally controlled pacing more effective (in some sense) than self-pacing?
5. Is one program equally effective for all students?
6. What are the effects of using different programing techniques (branching, intrinsic programing, vanishing) in various subject-matter areas?
7. What criteria are most appropriate in the evaluation of student learning?

These questions are but a few of the intriguing and complex problems facing investigators in the new field of programing material for teaching machines. Answers to these questions will help not only the educator but also the engineer who is concerned with writing adequate specifications for the construction of teaching machines.

Response Mode

The kind of response that should be given by a subject has been a controversial question in the teaching machine field. Pressey's original machine (1926) required the subject to press a lever corresponding to his choice of answer. The format of the answers was multiple-choice. Skinner (1958) emphasized the necessity of having the subject *compose* (construct) the response. Skinner states:

> One reason for this is that we want him to recall rather than recognize—to make a response as well as see that it is right. Another reason is that effective multiple-choice material must contain plausible wrong responses, which are out of place in the delicate process of "shaping" behavior because they strengthen unwanted forms (p. 2).

Coulson and Silberman (1960) investigated this question of multiple-choice vs. constructed response by using *simulated* teaching machines—human beings were used instead of automatic control mechanisms. Their results indicated that the multiple-choice response mode required significantly less time than the constructed response mode and that no significant difference was obtained between response modes on the criterion test. Further, they reported that no significant differences were obtained among the experimental groups on the multiple-choice criterion subtest or on the total (multiple-choice plus constructed response) criterion test. Fry (1959) has discussed this response-mode question along with other variables, and he has carried out extensive research concerning constructed vs. multiple-choice response modes. The results of his study favor the use of constructed response when recall is the objective of the learning.

In addition to the basic controversy (which needs much more investigation) between multiple-choice and constructed responses, there are several "variations on the theme" which are evident. Stephens (1953) has recommended that every wrong answer in a multiple-choice question appear as a correct choice for another item. He calls this program "inside alternatives." His data indicate that there was no difference between control and experimental groups on a criterion test using either nonsense syllables or Russian unless each right choice appeared as a wrong alternative for the three subsequent items. The use of prompts in general has been shown to be an effective technique in automated teaching (Cook, 1958; Cook & Kendler, 1956; Cook & Spitzer, 1960).

Using learning booklets, Goldbeck (1960) investigated the effect of response mode and learning material difficulty on automated instruction. The three response modes used were: overt response (the subject was required to construct a written response), covert response (the subject was permitted to think of a response), and implicit response (the subject read the response which was underlined). Goldbeck states:

> Learning efficiency scores, obtained by dividing quiz scores by learning time, showed that the implicit (reading) response condition produced significantly more efficient learning than the overt response condition. The covert response condition fell between the other conditions in learning efficiency (p. 25).

Concerning quiz-score results, the overt response group

> performed significantly poorer than the other response mode groups at the easy level of difficulty. Performance of the overt response group improved significantly at the intermediate difficulty level to the extent that it exceeded the performance of all other groups (pp. 25–26).

Goldbeck concludes that

> doubt is cast upon the assumption that the best learning is achieved by use of easy items and requiring written constructed responses (p. 26).

To the author's knowledge, the use of an oral response in conjunction with the Skinner teaching machine and its effect on learning rate and retention have not been reported in the literature. Furthermore, the importance of response mode as a function of reinforcement must be specified. Israel (1958) has suggested that natural and artificial reinforcement may affect the subjects' learning. A most comprehensive analysis of response-mode and feedback factors has been reported by Goldbeck and Briggs (1960).

The general area of reinforcement suggests problems related to the drop-out feature of teaching machines. Pressey's (1927) original machine dropped items after the correct answer had been given twice. Skinner's machines at the Harvard Psychological Laboratory also have the drop-out feature, although the commercially available machines based on Skinner's design do not incorporate this feature. With reference to a study carried out at Harvard, Holland (unpublished) reported significantly superior performance when the drop-out feature was used.

If items are dropped, the sequence of items is of course changed. How important is the sequence? If items should be dropped, by what criterion of learning can one justify omitting an item from the sequence? If items are not dropped and the criterion for the learning procedure is a complete run (i.e., once through the sequence without error), what is the effect upon retention? Being correct is positive reinforcement; thus, some items under these circumstances will receive a greater amount of positive reinforcement than others. What would be the effect of additional reinforcements with or without drop-out? Again, a plethora of problems and a paucity of answers!

Response time, another important variable, has been investigated by Briggs, Plashinski, and Jones (1955). Their study suggests that there is no difference between self-paced and automatically paced programs as determiners of response time. However, the problem of pacing for individual items is still a recent one and needs further research. Another aspect of response time—the distribution of practice—has been studied extensively since Ebbinghaus' investigation in 1885. For example, Holland (unpublished) states that in an experiment at Harvard "a few students completed all the disks in a small number of long sessions while others worked in many short sessions. . . . Apparently the way practice was distributed made little difference" (p. 4). Nevertheless, the distribution of practice, like the problem of pacing, is yet a subject of controversy, with most investigations favoring some form of distributed practice (Hovland, 1951).

The above section outlines briefly some of the major problems associated with the variables affecting response mode. Although some of the variables have already been investigated, these and others, together with their interactions, need further research.

Knowledge of Results

There are many peripheral problems related to teaching machines, one of which is the effect of immediate knowledge of results on learning. Angell (1949), using a multiple-choice punchboard technique, found that "learning is significantly enhanced by immediate knowledge of results." Briggs (1949), also using the Punchboard, confirmed these results. Bryan and Rigney (1956) noted superior performance when subjects were given knowledge of results, specifically, an explanation if the answer was incorrect. This last study was later expanded by Bryan, Rigney, and Van Horn (1957), who investigated differences between three kinds of explanation given for incorrect response. None of the three types of explanation proved to be superior in teaching the subjects. Because of their controvertible results, the above studies demonstrate that, although immediate knowledge of results appears to be effective in the learning process, this problem contains many facets which need more empirical data.

Motivation

One of the many reasons given for the effectiveness of teaching machines is that the student's motivation is increased. Psychologists and educators have realized for some time that the motivation variable ranks very high among those variables pertinent to learning. In 1958 and 1959, Holland surveyed the use of the teaching machine in classes at Harvard. He found that most students felt that they would have gotten less out of the course if the machines had not been used, that most students preferred to have machines used for part of the course, and finally that most students felt that the teaching machine was used by the instructor "to teach me as much as possible with a given expenditure of my time and effort." During a field tryout of the Subject-Matter Trainer in the Semiautomatic Ground Environment System, Mayer and Westfield (1958) observed that "motivation to work with the trainer is high." The supervisory as well as the operational personnel encouraged the use of this training technique.

Mager (1959) suggests that motivation and interest are a function of the percentage of correct responses. He observed that in two young subjects negative feelings for learning mathematics in the usual classroom situation did not transfer to learning mathematics by means of a teaching machine. The cause of this phenomenon is perhaps best explained by the subjects' statement that, because they were able to understand the programed material, it did not seem to be mathematics at all. This interesting relationship between comprehension and motivation needs further investigation.

Equipment

There are many inexpensive models of teaching machines which will soon hit the consumer market. For much of this equipment, there is very little experimental evidence which supports the various designs. As previously pointed out, Holland has collected data which support the efficiency of the drop-out feature in a teaching machine; yet commercial models presently available do not incorporate this feature, presumably because of its high cost. Generally, it seems that production is now and will continue to be out of phase with much of the research which has provided necessary teaching machine specifications. Moreover, because of their expense, it is likely that some very important features will be omitted in manufacture.

The methods of displaying programed material, another unexplored problem area, must be investigated so as to provide the design engineer with requirements based on empirical findings. The display problem is less acute, perhaps, with material for the elementary school than it is with programs designed to teach maintenance procedures and aspects of the biological sciences.

The use of computer controlled

teaching machines has been recommended by many authors (Coulson & Silberman, 1959; Skinner, 1958). Utilizing a central computer, with many programs capable of adapting to individual needs and of providing stimulus materials to 50 or more students simultaneously, is a feasible notion for large-scale training programs. With a computer, the display problem again becomes a major issue. Training in pattern recognition, information handling, and display interpretation are but a few appropriate areas which should be studied. The alternate modes of presentation become more extensive as computer capacity increases. In the case of certain kinds of subject matter, a computer generated, pictorial display of information may be a more effective presentation than other display techniques. Future research must solve these problems in equipment design.

Teaching Machines and Other Techniques

The use of automated teaching devices may be optimized, perhaps, if there is a proper balance between this technique and other compatible teaching methods. What percentage of a course should be machine taught? What subject matter is best suited to automated devices? If classroom courses were as carefully and thoughtfully programed as some of the programs currently being prepared for teaching machines, might some of the advantages of machines diminish? Perhaps some of the apparent advantages of teaching machines are no more than methods of illustrating correctable classroom techniques! It might well be that the instructor's enthusiasm and inspiration, a factor supposedly dominant in higher education, is vital in mastering a particular subject-matter area. Will creativity in certain students be harmed by extensive education via the machine? Again, consideration of the use of a teaching machine, the subject matter, the program, the level of education, and the techniques used in combination with the teaching machine provide a fertile field for experimentation. As of now, questions in this area remain unanswered. Silverman (1960a) has presented an excellent, detailed discussion of problems inherent in this new technology of automated teaching and the current trends in the field.

Problems of Application

The most obvious problems in the attempt to use automated teaching techniques have been outlined in the previous section. There is still much of the unknown associated with techniques, machines, programing, etc. to be eliminated before a direct solution to a particular training problem can be specified. Many alternatives exist, the best of which has not yet been determined. In addition to these voids, there is a serious lack of definition in the objectives of many training programs.

What is the objective of a particular automated course or program? From a pragmatic point of view, what are the criteria by which a specific educational program can be evaluated? For example, the objectives might range from the teaching of rote tasks to the presentation of more abstract material. Needless to say, the techniques for both teaching and evaluating learning could be substantially different in each case. The purpose of teaching, the objective of an educational program, must be initially defined. Only then will the concepts *learning* and *teaching* be meaningful in a particular context.

After definition, the next step is to determine what subject matter will provide the student with the necessary information. It is at this point that the major pitfall in education is likely to appear. Even though many training programs do not have a defined objective, their course content is nonetheless prescribed, and the text and/or materials used in previous, nonautomated courses become the prime source of material for an automated teaching program. To program an automated teaching machine with presently available materials might well result only in a more efficient method of teaching the wrong material!

The third step requires decisions in the selection of appropriate teaching techniques. Answers to questions involving programing, choice of teaching machine, learning procedures, pacing, and response modes are still not known.

The fourth and last step requires an evaluation of the selected automated teaching method in terms of the originally established objectives. Conventional methods of instruction should be compared with the innovative methods by means of a specific set of criteria, e.g., in terms of training time, job performance, retention of learned information, etc.

The questions confronting the researcher in teaching machine technology are one example of the broader questions of man-machine interrelation. Data pertinent to the principles of human engineering, the optimum man-machine interaction, the degree to which the machine can perform functions formerly allocated to man, and the appropriate allocation of functions between man and machine will be provided by a research program investigating teaching machines. Inadequate attention to any of the above-mentioned steps will result in failure to provide the needed answers in a field which may increase training effectiveness and reduce training costs.

REFERENCES

Angell, G. W. The effect of immediate knowledge of quiz results on final examination scores in freshman chemistry. *J. educ. Res.*, 1949, 42, 391–394.

Angell, G. W., & Troyer, M. E. A new self-scoring test device for improving instruction. *Sch. Soc.*, 1948, **67**, 84–85.

Beck, J. On some methods of programming. In E. H. Galanter (Ed.), *Automatic teaching: The state of the art.* New York: Wiley, 1959. Pp. 55–62.

Besnard, G. G., Briggs, L. J., Mursch, G. A., & Walker, E. S. Development of the subject-matter trainer. *USAF Personnel Train. Res. Cent. tech. Memo.*, 1955, No. ASPRL-TM-55-7.

Besnard, G. G., Briggs, L. J., & Walker, E. S. The improved subject-matter trainer. *USAF Personnel Train. Res. Cent. tech. Memo.*, 1955, No. ASPRL-TM-55-11.

Briggs, L. J. The development and appraisal of special procedures for superior students and an analysis of the effects of knowledge of results. *Abstr. Doctoral Dissertations, Ohio State U.*, 1949, No. 58.

Briggs, L. J., Plashinski, D., & Jones, D. L. Self-pacing versus automatic pacing of practice on the subject-matter trainer. *USAF Personnel Train. Res. Cent. lab. Note*, 1955, No. ASPRL-LN-55-8.

Bryan, G. L., & Rigney, J. W. An evaluation of a method for ship-board training in operations knowledge. *U. Sth. Calif. Electronics Personnel Res. Group tech. Rep.*, 1956, No. 18.

Bryan, G. L., & Rigney, J. W. Current trends in automated tutoring and their implications for naval technical training. *U. Sth. Calif. Dept. Psychol. tech. Rep.*, 1959, No. 29.

Bryan, G. L., Rigney, J. W., & Van Horn, C. An evaluation of three types of information for supplementing knowledge of results in a training technique. *U. Sth. Calif., Electronics Personnel Res. Group tech. Rep.*, 1957, No. 19.

Carr, W. J. Self-instructional devices: A review of current concepts. *USAF WADC tech. Rep.*, 1959, No. 59-503.

Cook, J. O. Supplementary report: Processes underlying learning a single paired-associate item. *J. exp. Psychol.*, 1958, **56**, 455.

Cook, J. O., & Kendler, T. S. A theoretical model to explain some paired-associate learning data. In G. Finch & F. Cameron (Eds.), *Symposium on Air Force human engineering, personnel, and training research.* Washington, D.C.: National Academy of Sciences–National Research Council, 1956. Pp. 90–98.

Cook, J. O., & Spitzer, M. E. Supplementary report: Prompting versus confirmation in paired-associate learning. *J. exp. Psychol.*, 1960, 59, 275–276.

Corrigan, R. E. Automated teaching methods. *Automated teach. Bull.*, 1959, 1(2), 23–30.

Coulson, J. E., & Silberman, H. F. Proposal for extension of automated teaching projects. *Sys. Develpm. Corp. field Note*, 1959.

Coulson, J. E., & Silberman, H. F. Effects of three variables in a teaching machine. *J. educ. Psychol.*, 1960, **51**, 135–143.

Crowder, N. A. *An automatic tutoring book on number systems.* Vol. 1. Timonium, Md.: Hoover Electronics Co., 1958.

Crowder, N. A. Automatic tutoring by means of intrinsic programming. In E. H. Galanter (Ed.), *Automatic teaching: The state of the art.* New York: Wiley, 1959. Pp. 109–116. (a)

Crowder, N. A. The concept of automatic tutoring. *USAF Personnel Train. Res. Cent. organizational Pap.*, 1959. (b)

Crowder, N. A. The "tutor." *J. Amer. Soc. Train. Dir.*, 1960, 14(5), 12–17.

Day, J. H. Teaching machines. *J. chem. Educ.*, 1959, **36**, 591–595.

Ferster, C. B., & Sapon, S. M. An application of recent developments in psychology to the teaching of German. *Harv. educ. Rev.*, 1958, 28, 58–69.

Fry, E. B. Teaching machine dichotomy: Skinner versus Pressey. Paper presented at American Psychological Association, Cincinnati, September 1959.

Fry, E. B., Bryan, G. L., & Rigney, J. W. Teaching machines: An annotated bibliography. *Audio-Visual commun. Rev.*, 1960, 8, Suppl. 1, 1–80.

Galanter, E. H. The ideal teacher. In E. H. Galanter (Ed.), *Automatic teaching: The state of the art.* New York: Wiley, 1959. Pp. 1–11.

Gavurin, E. I., & Donahue, V. M. Logical sequence and random sequence. *Automated teach. Bull.*, 1961, 1(4), 3–9.

Goldbeck, R. A. The effect of response mode and learning material difficulty on automated instruction. *Amer. Inst. Res. tech. Rep.*, 1960, No. AIR-328-60-IR-124.

Goldbeck, R. A., & Briggs, L. J. An analysis of response mode and feedback factors in automated instruction. *Amer. Inst. Res. tech. Rep.*, 1960, No. AIR-328-60-IR-133.

Holland, J. G. Teaching machines: An application of principles from the laboratory. In, Proceedings of the Educational Testing Service Invitational Conference, October 1959, *The impact of testing on the educational process.* Princeton, N. J.: Educational Testing Service, 1960.

Homme, L. E., & Glaser, R. Problems in programing verbal learning sequences. Paper presented in symposium on research issues in study of human learning raised by developments in automated teaching methods, American Psychological Association, Cincinnati, September 1959.

Hovland, C. I. Human learning and retention. In S. S. Stevens (Ed.), *Handbook of experimental psychology.* New York: Wiley, 1951. Pp. 613–689.

Israel, M. L. Skinnerian psychology and educational redesign. Paper read in symposium, American Psychological Association, Washington, D. C., September 1958.

Kendler, H. H. Teaching machines and psychological theory. In E. H. Galanter (Ed.), *Automatic teaching: The state of the art.* New York: Wiley, 1959. Pp. 177–185.

Little, J. K. Results of use of machines for testing and for drill upon learning in educational psychology. *J. exp. Educ.*, 1934, **3**, 45–49.

Mager, R. F. Preliminary studies in automated teaching. *IRE Trans. Educ.*, 1959, **E-2**, 104–107.

Mayer, S. R., & Westfield, R. L. A field tryout of a teaching machine for training in SAGE operations. *USAF Cambridge Res. Cent. tech. Memo.*, 1958, No. OAL-TM-58-16.

Peterson, J. C. A new device for teaching, testing, and research in learning. *Trans. Kans. Acad. Sci.*, 1930, **33**, 41–47.

Porter, D. Teaching machines. *Harv. Grad. Sch. Educ. Ass. Bull.*, 1958, 3(1), 1–5.

Pressey, S. L. A simple apparatus which gives tests and scores—and teaches. *Sch. Soc.*, 1926, 23, 373–376.

Pressey, S. L. A machine for automatic teaching of drill material. *Sch. Soc.*, 1927, 25, 549–552.

Pressey, S. L. A third and fourth contribution toward the coming "industrial revolution" in education. *Sch. Soc.*, 1932, **36**, 668–672.

Pressey, S. L. Development and appraisal of devices providing immediate automatic scoring of objective tests and concomitant self-instruction. *J. Psychol.*, 1950, 29, 417–447.

Pressey, S. L. Certain major psycho-educational issues appearing in the conference on teaching machines. In E. H. Galanter (Ed.), *Automatic teaching: The state of the art.* New York: Wiley, 1959. Pp. 187–198.

Ramo, S. A new technique of education. *Engng. sci. Mon.*, 1957, 21 (October), 17–22.

Rath, G. J., & Anderson, Nancy S. The IBM research center teaching machine project: I. The teaching of binary arithmetic. II. The simulation of a binary arithmetic teaching machine on the IBM 650. Paper presented at USAF Office of Scientific Research symposium on teaching machines, University of Pennsylvania, December 8–9, 1958.

Rath, G. J., Anderson, Nancy S., & Brainerd, R. C. The IBM research center teaching machine project. In E. H. Galanter (Ed.), *Automatic teaching: The state of the art.* New York: Wiley, 1959. Pp. 117–130.

Rothkopf, E. Z. A do-it-yourself kit for programmed instruction. *Teach. Coll. Rec.*, 1960, 62, 195–201.

Silberman, H. F. Introductory description of teaching machines (physical characteristics). Paper read in symposium on automated teaching, Western Psychological Association, San Diego, California, April 1959. (Abstract)

Silverman, R. E. Automated teaching: A review of theory and research. USN Training and Development Center, 1960. (ASTIA AD-241 283) (a)

Silverman, R. E. The use of context cues in teaching machines. USN Training and Development Center, 1960. (ASTIA AD-238 777) (b)

Skinner, B. F. The science of learning and the art of teaching. *Harv. educ. Rev.*, 1954, 24, 86–97.

Skinner, B. F. *Verbal behavior.* New York: Appleton-Century-Crofts, 1957.

Skinner, B. F. Teaching machines. *Science*, 1958, 128, 969–977.

Spence, K. W. The relation of learning theory to the technology of education. *Harv. educ. Rev.*, 1959, 29, 84–95.

Stephens, A. L. Certain special factors involved in the law of effect. *Abstr. Doctoral Dissertations, Ohio State U.*, 1953, No. 64.

Weimer, P. K. A proposed "automatic" teaching device. *IRE Trans. Educ.*, 1958, E-1, 51–53.

Zeaman, D. Skinner's theory of teaching machines. In E. H. Galanter (Ed.), *Automatic teaching: The state of the art.* New York: Wiley, 1959. Pp. 167–176.

(Early publication received March 5, 1961)

Psychological Bulletin
1961, Vol. 58, No. 5, 376–389

DEVELOPMENT OF RESEARCH ON THE PHYSIOLOGICAL MECHANISMS OF AUDITORY LOCALIZATION

MARK R. ROSENZWEIG
University of California, Berkeley

How does a listener determine the direction from which a sound comes? In the last century, physicists and physiologists usually gave a psychological explanation; they held that the listener makes a judgment by comparing the differing stimulation at the two ears. In the present century psychologists and physiologists have been seeking a physiological explanation; they are attempting to find where and how nerve impulses originating at the two ears interact in the brain. In this paper we will trace the development of hypotheses concerning the mechanisms of auditory localization. ("Localization" will signify here the perception of the *direction* of a sound source; perception of *distance* will not be considered.)

First, however, let us note briefly the importance of auditory localization in animal and human life. Some animals owe their livelihood to their ability to localize. Thus, certain types of bat catch their insect prey in the dark by echolocation, and certain moths, in turn, attempt to detect and avoid these bats by auditory clues (Griffin, 1958). Human beings also have considerable ability to localize sounds. Localization provides the basis for the detection of obstacles by blind people, an ability which long remained a mystery under the ambiguous designation of "facial vision of the blind." In the 1940s Cornell psychologists proved conclusively that this performance depends on localization of echoes reflected from the obstacles (Cotzin & Dallenbach, 1950; Supa, Cotzin, & Dallenbach, 1944). Seeing persons, too, benefit greatly from their ability to localize sounds. Localization makes it easier to listen to one signal or message in the presence of competing signals—a task which communication engineers refer to as "the cocktail party problem" (Cherry, 1957). Because sounds can be discriminated more readily when they are heard as coming from different directions, binaural hearing aids and stereophonic recordings of music are becoming steadily more popular.

Initial Work on the Mechanisms of Auditory Localization

The first person to investigate the nervous system in regard to localization seems to have been Louis Jurine, a naturalist, anatomist, and physician of Geneva. Lazaro Spallanzani of Pavia had demonstrated in 1793 that blinded bats could fly and avoid obstacles just as well as normal bats, but he could not imagine what sense then substituted for vision (1798). Jurine took up the question and decided that the solution must lie "at the tip of a scalpel." Noting the large size of the external ear of the bat, he went on to find that "a considerable neural apparatus" was devoted to hearing. Unfortunately, the published extract of his account (1798) does not give any fuller indication of Jurine's neuroanatomical findings. (The central connections of the auditory nerve seem not to have been discovered until almost a century later. In the bat the cochlear nucleus bulges out from the medulla right to the tip of the cochlea, and it is possible that Jurine took this for a large auditory nerve.) Following this lead, Jurine devised ingenious behavioral tests which provided clear evidence that the blinded bat guides itself by auditory clues.

Venturi and the Intensity Hypothesis

The first experiments on auditory localization by human observers seem to be those reported in two similar articles (1800a, 1800b) by Giovanni Battista Venturi, Professor of Physics at Modena and Pavia.[1] Venturi's experiments were similar to those that Lord Rayleigh performed independently 75 years later (1877). Venturi concluded that "The inequality of the two simultaneous sensations of the two ears informs us of the true direction of the sounds" (1800a, p. 386). Part of his evidence was that sounds coming from directly in front of the observer could not be distinguished from sounds coming directly from the rear, if the observer kept his head still. Venturi also concluded that a person with one deaf ear must turn his head to localize and will usually make errors in localizing sounds that are very brief. (As we shall see, certain recent experimenters have ignored the role of head movements in localization.)

Venturi noted that philosophers had attempted to explain the singleness of vision by the convergence of the two optic nerves, but he held that this was not the case for hearing:

> Since we distinguish the two simultaneous sensations of the two ears, and since their different intensities furnish us knowledge of the true direction of the sound, therefore one must conclude that the two sound impressions do not mix together inside the skull (1800a, p. 388).

Venturi furthermore concluded that the visual impressions of the two eyes do not mix, citing the phenomenon that was later to be called "retinal rivalry."

Venturi's intensity theory remained the dominant explanation for localization until early in the twentieth century. It was propounded, for example, by Magendie (1831) and Johannes Müller (1840). Magendie offered it as something evident and did not credit its discovery to anyone. Müller claimed that perception of direction of sound "is an act of judgment which founds it on experience previously acquired. . . . The only true guide for this inference is the more intense action of the sound upon one than upon the other ear" (p. 479). He further noted that when a sound comes from directly ahead or behind, it falls equally upon the two ears and is then impossible to localize; this demonstration he ascribed to "Venturini" (sic).

Progress in the Second Half of the Nineteenth Century

The first person to abandon the judgmental interpretation of localiza-

[1] This work of Venturi seems to have set the style for most of the research on auditory localization during the nineteenth century. Yet, while some of his findings soon became common knowledge, it was forgotten who had discovered them. Thus Klemm, in his detailed history of auditory localization (1914), mentioned Venturi in a single sentence and only in regard to effects of head movements. Pierce (1901) and Boring (1942) did not mention Venturi in their historical discussions of localization; both ascribed the first experiment on localization to E. H. Weber in 1848.

Venturi's first publication on this subject is even earlier than those indicated above. A paper of 1796 in French gives almost the same material as that of the German articles of 1800 (see Venturi, 1796).

In 1801 Venturi published a report in Italian on this research, appending it to the second edition of his book on physical research on color (see Venturi, 1801). The contents of this version are similar to those of the German articles. Thus it appears that Venturi made repeated attempts to secure a wide public for this research.

tion seems to have been S. Scott Alison, an English physician. He had invented the "differential stethophone" which consisted simply of two stethoscopes, one for each ear. After using this instrument, Alison reported that a sound is restricted to the ear that receives it in greater intensity and is suppressed in the other ear. This, he remarked,

> holds apparently in virtue of a law seemingly established for the purpose of enabling man and the lower animals to determine the direction of the same sound, with more accuracy than could be done had a judgment to be formed between the intensity of two similar sensations in the two ears respectively (1858, pp. 388–389).

Alison was considerably ahead of his time in this formulation but his work seems to have had little influence.

Sylvanus P. Thompson wondered about the basis of binaural beats, heard when he connected two slightly mistuned forks, one to each ear (1877). (Lord Rayleigh was to observe later—1907—as Dove—1857—had earlier, that the sound also changed location while beating.) Thompson rejected the hypothesis that bone conduction could account for binaural beats. Noting that the auditory nerves do not decussate as the optic nerves do, he concluded "that any means of comparison which may exist in the nerve systems of the ears exists deep-seated in the actual structure of the brain" (1877, p. 276). In his next paper he again noted that it is problematical where the sensations from the two ears "blend," and he remarked, "This point deserves the attention of anatomists and physiologists" (1878, p. 389).

Tracing the Afferent Auditory Pathways

Physiologists and anatomists had, in fact, already turned their attention to the localization of sensory processes in the brain. The study of cerebral localization of function had recently been given new impetus by the introduction of a new method—precise electrical stimulation, introduced by Fritsch and Hitzig in 1870. The revolutionary results obtained by this method led also to renewed interest in experiments involving precise ablation, as a check on the electrical experiments. David Ferrier, Professor of Neuropathology in London, began extensive mapping of the brain of several species in 1873, using electrical and surgical techniques (1890). In the superior temporal convolution of monkeys and its homologues in other species, electrical stimulation produced the same reaction as if a shrill sound had been made in the contralateral ear. The animal pricked up or retracted the ear and often moved its head or eyes to that side. (Much later the same experiment was to be performed in conscious surgical patients—Penfield & Rasmussen, 1950. In most cases, the human subjects reported hearing sounds on the side contralateral to the stimulated hemisphere; some sounds were heard "bilaterally"; no sounds were heard ipsilaterally.) Ferrier also claimed that ablation of the auditory cortex of both hemispheres made monkeys inattentive to sound. Heschl in 1878 succeeded in tracing the auditory tracts to the superior temporal convolution, and the auditory area is often given his name.

Luigi Luciani of Florence used the ablation technique extensively in cortical mapping (Luciani, 1884; Luciani & Seppilli, 1886; Luciani & Tamburini, 1879). One behavioral test devised by Luciani has recently been reintroduced by Riss (1959). In this test, bits of food were thrown to the floor near the blindfolded animal, and the accuracy of its reactions to the sounds was noted. Luciani confirmed that the auditory area of the cortex is located posteriorly in the temporal lobe. He found that

> each ear [has] connections with both auditory spheres, but chiefly with that of the opposite side. In fact, every unilateral extirpation of sufficient extent in the province of the auditory sphere causes a bilateral disorder of hearing, more marked on the opposite side . . . (1884, p. 155).

Thus, after an ablation in the right hemisphere, these results were reported by Luciani and Seppilli:

> Hearing is affected on both sides, but more at the left than at the right ear. The animal shows that it hears the sound of pieces of food falling to the left, but it mistakes the direction and turns to the other side. At the right ear this does not happen (1886, p. 79).

The effects of unilateral lesions generally disappeared within a few weeks; they persisted longer the larger the lesion. None of the lesions covered the whole auditory area as it is now defined, so it cannot be told from these studies whether a complete unilateral lesion would have led to some permanent impairment of localization. Bilateral lesions of the auditory areas were found to produce permanent perceptual impairment, as this example indicates:

> When called suddenly, the dog reveals through its movement that it is not deaf; but it does not follow and does not turn its head toward the sound, but, in fact, often even turns to the other side; in short, it seems not to understand what it hears and not to perceive the direction of sounds (psychic deafness) (1886, p. 119).

Luciani concluded that in the auditory system, just as in the visual system

> we must distinguish a crossed and a direct fasciculus; the former consisted of a much larger number of fibers than the latter. Neither of these fasciculi possesses any uniform relation with distinct segments of their respective cortical spheres, but their fibers irradiate themselves throughout the area of these centres (1884, p. 155).

Histological degeneration studies of Baginsky, Flechsig, and von Monakow soon revealed rather completely the course of the auditory pathway, crossing in part in the brainstem and proceeding by way of the inferior colliculus and medial geniculate body to the cortex of the temporal lobe (Ferrier, 1890).

After these early achievements, progress not only lagged but some of the findings were even forgotten by workers in the field. Thus, for example, it was taken as surprising in 1928 when removal of one cerebral hemisphere of a patient did not destroy hearing in the opposite ear (Bunch, 1928).

THE EARLY TWENTIETH CENTURY: THE PHASE HYPOTHESIS

Although the role of dichotic phase differences in auditory localization had been shown by Dove (1857), Thompson (1877), and others in the nineteenth century, the phase hypothesis was firmly established only in the twentieth century. Lord Rayleigh had considered this hypothesis previously, but his advocacy of it in 1907 convinced others. Rayleigh showed that while dichotic intensity differences permit localization of high frequency sounds, it is dichotic phase differences that permit localization of low frequency sounds. This belated recognition of the dual basis of localization might have been expected to embarrass supporters of the judgmental position. In fact, Rayleigh did remark, "Perhaps it is not to be expected that we should recognize intuitively the very different basis upon which our judgment rests in the two cases" (1907, p. 203). Nevertheless he never abandoned the judgmental interpretation of localization.

Bowlker (1908), who experimented on the role of phase differences, speculated briefly about neural interaction:

> we may suppose that the transmission of sound impulses through some specialized part of the auditory apparatus or brain takes a definite time from each ear, and that the point where the impulses meet is the focus that gives rise to the sensation of a sound image (p. 327).

From the vantage point of the present, this seems to anticipate Jeffress' hypothesis of 1948, but Bowlker's suggestion is so terse that we cannot be sure.

The success of the Stenger test for unilateral auditory malingering (1907) could have been taken as evidence against the judgmental approach to localization, but it does not seem to have been. The test, recently termed "the most reliable and effective of all malingering tests" (Watson & Tolan, 1949), works in the following way: A subject who simulates deafness of one ear will report hearing a tone that is delivered only to his other ear. When the tone is next delivered to both ears, and more intensely to the ear whose deafness is feigned, the subject will hear it only at this supposedly deaf ear. The malingerer will therefore give himself away by reporting that he does not hear the sound, in spite of the fact that it is present in audible intensity at the admittedly good ear. Thus, it was clear to Stenger (and it should have been to all users of his test) that the listener hears only a single sound and does not compare separate sensations arising from the two ears. This advance in clinical testing had no apparent influence on the development of thinking about auditory localization.

The Time Hypothesis

The hypothesis that dichotic time provides a basis for localization seems first to have been proposed seriously by Mallock (1908) and first to have been demonstrated experimentally by Aggazzotti (1911). It was brought to wide attention in several publications at the end of the first World War (Klemm, 1918, 1920; Piéron, 1922; von Hornbostel & Wertheimer, 1920). Dichotic stimuli separated by as little as 30 microseconds were shown to be perceived toward the side of the prior component. The time hypothesis is incompatible with the judgmental approach, for the dichotic time intervals which provide for localization lie under the threshold of fusion of successive auditory stimuli. That is, dichotic stimuli with a time interval less than 2 milliseconds give rise to perception of a single localized auditory event; there are not two perceptual events that can be compared in order to judge localization. Perhaps the first to recognize the incompatibility of the time hypothesis and the judgmental approach were Kreidl and Gatscher (1923). Their conclusion was to reject the time hypothesis! Since they showed that stimuli must be separated by about 20 milliseconds to be judged as successive, they denied that smaller intervals could have any effect in perception. von Hornbostel (1926) showed the fallacy of this argument. Furthermore, any observer who attempted to test the time hypothesis could verify it. The success of the time hypothesis thus helped to overcome the judgmental approach and to clear the way for work on the physiological mechanisms of localization.

Hypothesized Central Mechanisms

After the role of time differences was demonstrated, several further hypotheses about the mechanisms of localization were soon proposed. von

Hornbostel (1926) suggested that intensity differences were converted into time differences in the auditory system, a stronger stimulus evoking neural responses with less latency than a weaker stimulus. Kemp and Robinson (1937b) were able to demonstrate that the latency of auditory impulses does decrease with intensity, but only within 40 db. of threshold. Stevens and Davis in their book, *Hearing, Its Psychology and Physiology* (1938), mentioned the work of Kemp and Robinson and concluded that the effect of intensity differences cannot result solely from changes in latency. Their reason was that changes in the binaural intensity ratio can shift the location of intense tones. In fact, the change of ratio had been found to be smallest when the tone was about 80 db. above threshold (Upton, 1936); at this level there is no longer a change in latency, according to Kemp and Robinson's results. (This point, we may note, is all that Stevens and Davis had to say about the physiological correlates of auditory localization.) Later research (e.g., Pestalozza & Davis, 1956) has shown that latency continues to decrease with intensity up to at least 70 db. above threshold; this gives new support to von Hornbostel's hypothesis.

Boring (1926) suggested that the locus of cortical excitation might be the physiological correlate of localization. He hypothesized that the ears project, in each cerebral hemisphere, to cortical areas that are not coincident but which overlap. If one ear was stimulated either earlier or more strongly than the other, then the cortical excitation would be located mainly in the projection area of that ear.

Trimble (1928) proposed a vague central hypothesis in which the interaural differences are transmitted to the cortex where localization occurs.

> The directional localization of a sound source, under ordinary conditions of hearing, depends upon the configurational nature of the cortical effects that correspond to the physical "difference-pattern" at the ears (p. 523).

von Békésy (1930) proposed a rather detailed schema. He pictured a region of cells where the auditory tracts from the two ears join. Auditory localization would depend upon the proportions of the region that each side excited. Both greater intensity and prior arrival would favor the ear so stimulated.

Woodworth described a possible mechanism in the discussion of auditory localization in his text, *Experimental Psychology* (1938):

> It must be a unitary mechanism capable of turning the head in either direction and responsive to nerve currents from both ears. When the currents arrive from both ears, but more from one ear, that side has the advantage. When the current from one ear arrives at the central mechanism ahead of the other and gets in its work first, it has the advantage (p. 533).

Jeffress (1948) suggested a neural "mechanism for the representation of a time difference as place" in the auditory system. He pictured a center where tracts from both ears make common synaptic connections. Within this center there are places where the conduction time is slightly longer from one ear than from the other. If the two ears are stimulated simultaneously, the impulses meet and summate at the locus where the conduction times from both sides are equal. If one ear is stimulated before the other, then the impulses meet at a different locus—a locus where the difference in conduction times compensates for the dichotic time difference. Intensity difference is translated into time difference and this is then handled in the same way. Jeffress ventured that this mechanism might be located in the medial geniculate, relying on the electrophysiological evidence of Kemp and Robinson (1937a) that no binaural interaction could be found at the lateral lemniscus. Later Jeffress disavowed this location and suggested that if his hypothesized mechanism exists, it exists in the accessory nucleus of the superior olive (1958).

Localization Investigated by Physiological Techniques

While these hypothetical mechanisms were being proposed, further physiological findings were being obtained by both ablation and electrophysiological techniques.

Information from Ablation Studies

Pavlov (1927) reported a finding of Bikov that a dog could not learn to discriminate between right and left positions of a sound source after the corpus callosum had been transected (p. 150). Girden (1940) found, on the contrary, that dogs retained a learned right-left discrimination of a tone or bell after transection of the corpus callosum. Girden's animals were trained to respond by flexing a leg when the sound came from one side and not to respond when it came from the other. Each sound lasted for 2 seconds. They also retained the discrimination after hemidecortication, losing it only after complete bilateral ablation of the auditory cortex. Even after bilateral ablation, the dog could still orient itself to sound, but it could not be retrained to show the conditional discrimination.

ten Cate (1934) reported that decorticate cats could orient to a sound stimulus. He used various sounds, all of them lasting for 15 to 20 seconds. The physiologists Bard and Rioch (1937) also reported that decorticate cats could localize sounds accurately, but they did not offer any quantitative observations.

Measurement of impaired ability to localize by cats with bilateral ablations of the auditory cortex was furnished by Neff and his collaborators (see Neff & Diamond, 1958, for a history of several stages of this research). Cats were trained to go to the one of two boxes behind which a buzzer sounded. Intact animals could do this when the boxes were only 5 degrees apart, the angle being measured from the point at which the cats were released into the test area. Animals with complete bilateral destruction of the auditory cortices could discriminate only when the boxes were 40 degrees apart. Three different hypotheses were advanced, any of which might account for the observed results: (*a*) "An intact auditory cortex is necessary in order that the relationship between auditory signal and food reward may be learned" (Neff, Fisher, Diamond, & Yela, 1956, p. 510). Further experimentation refuted this hypothesis, since cats with bilateral ablation of auditory areas could learn to open a single door when a buzzer sounded. (*b*) "An intact auditory cortex is essential for maintenance of attention to an auditory signal." (*c*) "An intact auditory cortex is necessary for accurate localization of sound in space" (1956, p. 511). Neff and Diamond (1958) also reported preliminary results indicating that ability to localize in their tests

> is not affected by section of the corpus callosum, is affected very little if at all by section of the commissure of the inferior colliculus, but is severely affected by section of the trapezoid body (p. 108).

Riss (1959) noted that auditory signals of relatively long duration had been employed in the studies of ten Cate and Neff. He raised the question whether *binaural* localization was actually tested in their work, since it is well known that monaural localization is possible if head movements can be performed while the sound continues. (It will be remembered that Venturi had made this point in 1800.) Riss pointed out that animals of Bard and Rioch oriented to the stimulus slowly, using noticeable head and ear movements. Riss therefore employed both very brief sounds and sounds lasting as long as 30 seconds in experiments with cats. For brief sounds, bits of food were thrown down beside the cat; this is the method that Luciani had used 75 years previously, although Riss evidently did not know of his work. The results replicated and extended those of Luciani. Cats with control lesions were successful with both types of signal. Cats with bilateral ablation of the auditory areas "showed evidence of being unable to orient to brief sounds but were partially successful in seeking out the region of the sound if the sound was prolonged" (p. 383). Tests revealed that these animals could maintain attention to sound. Riss therefore concluded "that the auditory cortex is necessary for localizing the instantaneous position of a sound" (p. 383).

Information from Electrophysiological Studies

Wherever in the brain the tracts from the two ears converge functionally, it should be possible to find interaction between the electrophysiological responses. Kemp and Robinson (1937a) recorded from the brain stem of the anaesthetized cat while presenting monaural or binaural tones or clicks. They interpreted their results as showing no binaural interaction at the level of the lateral lemniscus, arguing "against the convergence of the tracts from the two ears in the cochlear nuclei or superior olivary complex" (p. 322).

The relative representation of the two ears at the auditory cortex was next investigated in several studies, beginning with that of Bremer and Dow on the cat (1939). Bremer and Dow reported that the response to stimulation of either ear was greater at the contralateral cortex. In contrast to this, Woolsey and Walzl (1942) concluded:

> each cochlea is bilaterally represented in the primary projection area, and impulses from corresponding points of each cochlea terminate in common areas of each hemisphere. Ipsilateral and contralateral representations for each cochlea appear to be equal (p. 341).

Tunturi (1944, 1946), working on the dog, reported that each cochlea is represented over the whole extent of both auditory cortices and that each is represented slightly more strongly at the contralateral cortex. Rosenzweig (1951) measured the amplitudes of series of ipsilateral and contralateral cortical responses at 49 electrode placements in five cats. Statistical tests demonstrated that the contralateral response was significantly larger than the ipsilateral at 28 locations; neither was significantly larger at 20 locations, and the ipsilateral response was significantly larger at only 1 location. Thus, the experimenter could tell which ear had been stimulated, either by comparing the amplitudes of responses at the two hemispheres or by inspecting the pattern of amplitudes within a single hemisphere. The stronger representation of each ear in the contralateral hemisphere confirmed the results that Luciani had obtained much earlier in his ablation studies.

Tunturi (1946) had also obtained

evidence of binaural interaction at the cortex. Since his dichotic stimuli were usually separated by intervals of several milliseconds, it remained to determine whether dichotic intervals of a fraction of a millisecond could be preserved in afferent transmission all the way up to the cortex and used there for binaural interaction. Although transmission from the cochlea to the cortex requires about 10 milliseconds, dichotic time intervals of one-tenth of a millisecond were found to affect response amplitudes significantly (Rosenzweig, 1954; Rosenzweig & Rosenblith, 1950). The relation between amplitudes of simultaneous responses at the auditory areas of the two hemispheres was shown to correlate with auditory localization. "At either hemisphere the amplitude of the summated response is larger when the contralateral ear receives the prior stimulus" (Rosenzweig & Rosenblith, 1950, p. 879). "The cortical events were found to parallel in several respects the perceptual phenomena which occur under the same stimulus conditions" (Rosenzweig, 1954, p. 275). Bremer (1952) arrived independently at the hypothesis that the relation between the amplitudes of responses at the two hemispheres is the cerebral index to auditory localization.

Tests were then made for binaural interaction at lower levels of the auditory system. Ades and Brookhart had suggested "that the inferior colliculus with its strong commissural connections and connections to afferent mechanisms may be the principal device responsible for localization" (1950, p. 203). Interaction was found at the inferior colliculus (Coleman, 1953; Rosenzweig & Wyers, 1955), but the importance attributed to the commissural connections by Ades and Brookhart was thrown into doubt by the following observations: Transecting the commissure of the colliculi does not affect interaction (Rosenzweig & Wyers, 1955); moreover, this transection does not impair auditory localization (Neff & Diamond, 1958). At the colliculi, as at the cortex, recordings made with macroelectrodes show that stimulation of the contralateral ear evokes responses of greater amplitude than does stimulation of the ipsilateral ear (Rosenzweig & Wyers, 1955). Using microelectrodes, Erulkar (1957) found that of 89 single units tested with click stimuli, 23 responded only to stimulation of the contralateral ear, 11 responded only to stimulation of the ipsilateral ear, and 55 responded to stimulation of either ear. For most units that responded to either ear, the latency of response was nevertheless shorter for contralateral than for ipsilateral stimulation. Furthermore, latency showed rather regular changes as the position of the stimulus was moved around the head of the experimental animal (Erulkar, 1959).

Proceeding further down the auditory system, evidence of binaural interaction was also found at the lateral lemniscus (Rosenzweig & Amon, 1955; Rosenzweig & Sutton, 1958). This finding refuted the conclusion of Kemp and Robinson (1937a) that tracts from the two ears do not converge before the colliculi. Kemp and Robinson had found no signs of interaction when they used stimuli dichotic in time, but they gave no details, not even the time intervals employed. Rosenzweig and Sutton, on the contrary, presented measures of the reduction in amplitude of the response to the second stimulus as a function of the dichotic interval.

The lowest level at which binaural interaction occurs may be the superior olivary nuclei. Stotler (1953) reported these findings concerning the anatomy of the olivary nuclei:

> The cells of the medial superior olivary nucleus receive afferent terminals in the form of boutons from both cochlear nuclei. The afferent fibers from the homolateral cochlear nucleus terminate on the lateral pole of the cell, while those from the contralateral cochlear nucleus end in relation to the medial pole. The axons of the medial nucleus enter the homolateral lateral lemiscus (p. 420).

Thus the cells of this nucleus seem ideally situated to integrate information from the two ears. Subsequent electrical recording has in fact shown that some of the superior olivary nuclei show differences in responses depending upon time differences in stimulation of the two ears (Galambos, Schwartzkopff, & Rupert, 1959).

> Units in n. accessorius proved to be exquisitely sensitive to whether the right ear or the left was stimulated first by paired clicks; the unique physiological and anatomic characteristics of these cells seem relevant to the binaural sound localization problem (p. 527).

Thus the primary neural interactions basic to localization may occur low in the afferent pathways.

It had been suggested in a preliminary report that interaction might occur at as low a level as the cochlea itself, impulses being transmitted from one cochlea to the other (Galambos, Rosenblith, & Rosenzweig, 1950). The transmission appeared to require about 1 millisecond, and some evidence of interaction was obtained when a click at one ear preceded that at the other ear by 1.25 milliseconds. In a later study no evidence of interaural interaction was obtained, using a dichotic interval of 3.6 milliseconds (Rosenblith & Rosenzweig, 1951). It now appears that the latter interval may have been too great, since interaction at the lateral lemniscus can be observed clearly only if the dichotic interval is less than about 3 milliseconds (Rosenzweig & Sutton, 1958). More telling evidence against the occurrence of interaural interaction in the cochlea is the failure to find it in the cochlear nucleus under conditions where significant interaction was found in the olivary nucleus and in the lateral lemniscus (Rosenzweig & Amon, 1955). Moreover, as Walsh (1957) has pointed out, interaction at the cochlear level, even if it occurred, could not mediate localization. The time required for transmission from one cochlea to the other seemed to be of the order of 1 millisecond, while localization is obtained with dichotic intervals of one-tenth of a millisecond or less. Thus the lowest level of the auditory system at which binaural interaction has been demonstrated is that of the olivary nuclei. The ablation studies nevertheless suggest that the auditory cortex must be involved if neural interaction in the brain stem is to eventuate in behavioral discrimination of location of sound sources.

Localization with Only One Cerebral Hemisphere?

Walsh (1957) tested 22 patients with various cerebral defects to find how well they could localize sounds. In several of these cases the auditory area of one hemisphere was probably impaired, and in one case the right cerebral hemisphere had been removed. Most of these patients, including the case of hemispherectomy, were able to localize on the basis of time differences between clicks presented at the two ears by means of earphones. With most subjects only a few dichotic intervals were employed, and thresholds were not determined. The intervals were usually of the order of 300 to 500 microseconds—several times the threshold intervals commonly reported for normal subjects. The hemispherectomized patient could localize with intervals of 410 and 190 microseconds but not with an interval of 125 milliseconds. Walsh concluded, "The sensitivity to binaural time differences is retained after the loss of the auditory cortex on one side" (p. 248). Since thresholds were not determined, it cannot be concluded that there was no impairment of localization with loss of the auditory cortex of one hemisphere.

Accurate thresholds for dichotic stimulation were determined for brain injured subjects in a study reported briefly by Teuber and Diamond (1956). Twenty patients with penetrating brain injuries, 14 of them unilateral, were compared with 10 control subjects who had injuries of peripheral nerves. The brain injured subjects, compared to the controls, required a significantly larger dichotic interval to shift a click from the median plane; the thresholds were 225 and 105 microseconds, respectively. Similarly, the difference in intensities at the two ears necessary to shift a click from the center location was significantly larger for brain injured than for control subjects; the thresholds were 11 and 5 db., respectively. Subjects with unilateral lesions in the right hemisphere required greater intensity on the left than on the right side in order to judge the sound at the midline, and conversely for subjects with left unilateral lesions. (This is similar to the results found by Luciani and Seppilli—1886—with unilateral ablations.) No such directional characteristic was found in the impairment of judgments involving dichotic time. The subjects who were impaired with respect to thresholds for dichotic time were not necessarily impaired with respect to thresholds for dichotic intensity, and conversely. This suggested that the neural mechanisms for localization based on these two cues might not be identical.

Coleman (1959) recorded electrical responses from several positions on the auditory cortex of anaesthetized cats while either moving a click source around the animal's head or varying dichotic time and intensity of clicks produced at the two ears. The relative amplitudes of responses at different electrode positions varied with the location of the sound source or with the dichotic conditions. Some points gave larger responses to contralateral and some to ipsilateral stimuli. "These data suggest that angular location of auditory stimuli may be represented in the auditory cortex of one hemisphere by means of a place principle" (p. 40).

These observations are hard to reconcile with the results of experiments in which electrical stimuli were applied to the auditory area of one cerebral hemisphere (Ferrier, 1890; Penfield & Rasmussen, 1950). It will be remembered that human subjects, under these conditions, usually hear a sound contralaterally to the side stimulated; sometimes they hear the sound on both sides, but never do they hear it ipsilaterally to the stimulation. This suggests that it is not possible to excite points in the right hemisphere that represent locations in space to the right of the listener, nor points in the left hemisphere that represent locations to the left of the listener.

On the basis of his findings Coleman was inclined to reject what he termed "the bilateral ratio theory" of Rosenzweig. No sure conclusion concerning ability to localize can be drawn from Coleman's observations, for the position of a sound source is also represented at the olivary nucleus, yet a cat cannot localize accurately using only the lower brain centers. Should Walsh's conclusion be substantiated that the cortex of one hemisphere suffices for normal auditory localization, then Coleman's finding may take on increased importance.

Formulation of a comprehensive hypothesis about the mechanisms of binaural localization requires an answer to the question whether the cortex of a single hemisphere is sufficient for normally accurate localization. This question may be resolved by precise determination of the capacity to localize in patients or animals with complete unilateral destruction of the auditory cortex.

Summary and Conclusion

From the work of Venturi in 1800 until about 1920, the perception of location of a sound source was generally considered to be a judgment arrived at by comparing differences in the stimulation at the two ears. While Venturi showed that monaural localization was possible, if the listener could move his head during the presentation of the sound, the chief interest has always been in binaural localization. Over most of this period, dichotic difference in intensity was considered to be the only or the main stimulus basis for localization. Comparison of dichotic intensities seemed to be a plausible explanation for localization, even though Alison pointed out in 1858 that a sound nearer one ear is heard at that side only and seems to be suppressed in the other ear. While the judgmental interpretation and the intensity hypothesis reigned, there was little incentive and little effort to work out the physiological mechanisms of auditory localization. This was true in spite of noteworthy advances in knowledge of the anatomy of the afferent auditory system, beginning in the 1870s.

The establishment of the dichotic time hypothesis at the end of the first world war was quickly followed by abandonment of the judgmental position. This position could no longer be maintained when it was realized that the time intervals on which localization is based are too small to be perceived as intervals; only a single localized sound is heard. Psychologists soon proposed a number of speculative mechanisms for localization, involving interaction of neural impulses converging from the two ears upon some central locus. During the last 25 years a number of experimenters have brought ablation and electrophysiological techniques to bear on the problems of localization. They have recently shown that the cortex is required for binaural localization, although neural interaction first occurs low in the brain stem. Some evidence suggests that the cortex of a single hemisphere may be sufficient to permit localization. A completely satisfactory hypothesis of the mechanisms of binaural auditory localization, including both cortical and subcortical components, is yet to be presented.

References

Ades, H. W., & Brookhart, J. M. The central auditory pathway. *J. Neurophysiol.*, 1950, **13**, 189–205.

Aggazzotti, A. Sul più piccolo intervallo di tempo percettibile nei processi psichici. *Arch. Fisiol.*, 1911, **9**, 523–574.

Alison, S. S. On the differential stethophone, and some new phenomena observed by it. *Phil. Mag.*, 1858, **16**, 385–395.

Bard, P., & Rioch, D. E. A study of four cats deprived of neocortex and additional portions of the forebrain. *Bull. Johns Hopkins Hosp.*, 1937, **60**, 73–147.

Boring, E. G. Auditory theory with special reference to intensity, volume and localization. *Amer. J. Psychol.*, 1926, **37**, 157–188.

Boring, E. G. *Sensation and perception in the history of experimental psychology.* New York: Appleton-Century, 1942.

Bowlker, T. J. On the factors serving to determine the direction of sound. *Phil. Mag.*, 1908, **15**, 318–332.

Bremer, F. Les aires auditives de l'écorce cérébrale. In, *La surdité: Sa mésure et sa correction.* Paris: Maloine, 1952.

Bremer, F., & Dow, R. S. The acoustic area of the cerebral cortex in the cat. *J. Neurophysiol.*, 1939, **2**, 308–318.

Bunch, C. C. Auditory acuity after removal of the entire right cerebral hemisphere. *J. Amer. Med. Ass.*, 1928, **90**, 2102.

Cherry, C. *On human communication.* New York: Wiley, 1957.

Coleman, P. D. An electrophysiological study of the interaction between responses to successive clicks in the inferior colliculus of the cat. Unpublished doctoral dissertation, University of Rochester, 1953.

Coleman, P. D. Cortical correlates of auditory localization. *Science*, 1959, **130**, 39–40.

Cotzin, M., & Dallenbach, K. M. Facial vision: The role of pitch and loudness in the perception of obstacles by the blind. *Amer. J. Psychol.*, 1950, **63**, 485–515.

Dove, H. W. Eine akustische Interferenz. *Ann. Phys. Chem.*, 1857, **101**, 492–494.

Erulkar, S. D. The responses of the inferior colliculus of the cat to binaural stimulation. *J. Physiol.*, 1957, **138**, 34–35. (Abstract)

Erulkar, S. D. The responses of single units of the inferior colliculus of the cat to acoustic stimulation. *Proc. Roy. Soc. Lond., Ser. B*, 1959, **150**, 336–355.

Ferrier, D. *The Croonian lectures on cerebral localization.* London: Smith, Elder, 1890.

Galambos, R., Rosenblith, W. A., & Rosenzweig, M. R. Physiological evidence for a cochleo-cochlear pathway in the cat. *Experientia*, 1950, **6**, 438–440.

Galambos, R., Schwartzkopff, J., & Rupert, A. Microelectrode study of superior olivary nuclei. *Amer. J. Physiol.*, 1959, **197**, 527–536.

Girden, E. The role of the auditory area of the cortex. *Amer. J. Psychol.*, 1940, **53**, 371–383.

Griffin, D. R. *Listening in the dark.* New Haven: Yale Univer. Press, 1958.

Heschl, R. L. *Ueber die vorderquere Schläfenwindung des menschlichen Grosshirns.* Vienna: Braumüller, 1878.

Jeffress, L. A. A place theory of sound localization. *J. comp. physiol. Psychol.*, 1948, **41**, 35–39.

Jeffress, L. A. Medial geniculate body: A disavowal. *J. Acoust. Soc. Amer.*, 1958, **30**, 802–803.

Jurine, L. Extrait des expériences de Jurine sur les chauves-souris qu'on a privé de la vue. *J. Phys.*, 1798, **46**, 145–148.

Kemp, E. H., & Robinson, E. H. Electric responses of the brain stem to bilateral auditory stimulation. *Amer. J. Physiol.*, 1937, **120**, 316–322. (a)

Kemp, E. H., & Robinson, E. H. Electric responses of the brain stem to unilateral auditory stimulation. *Amer. J. Physiol.*, 1937, **120**, 304–315. (b)

Klemm, O. Sammelreferat über die Lokalisation von Schallreizen. *Kongr. exp. Psychol.*, 1914, Ber. 6, 169–258.

Klemm, O. Untersuchungen über die Lokalisation von Schallreizen: III. Mittelung: Ueber den Anteil des beidohrigen Hörens. *Arch. ges. Psychol.*, 1918, **38**, 71–114.

Klemm, O. Ueber den Einfluss des binauralen Zeitunterschiedes auf die Lokalisation. *Arch. ges. Psychol.*, 1920, **40**, 117–146.

Kreidl, A., & Gatscher, S. Ueber die dichotische Zeitschwelle. *Pflüg. Arch. ges. Physiol.*, 1923, **200**, 366–373.

Luciani, L. On the sensorial localisations in the cortex cerebri. *Brain*, 1884, **7**, 145–160.

Luciani, L., & Seppilli, G. *Die Funktions-Localisation auf der Grosshirnrinde.* (Trans. by M. O. Fraenkel) Leipzig: Denicke's, 1886.

Luciani, L., & Tamburini, A. *Ricerche sperimentali sui centri psico-sensori corticali.* Reggio Emilia: Calderini, 1879.

Magendie, F. *An elementary compendium of physiology.* (4th ed.) (Trans. by E. Milligan) Edinburgh: Carfrae, 1831.

Mallock, A. Note on the sensibility of the ear to the direction of explosive sounds. *Proc. Roy. Soc. Lond., Ser. A*, 1908, **80**, 110–112.

Müller, J. *Handbuch der Physiologie des Menschen.* Vol. 2. Coblenz: Hölscher, 1840.

Neff, W. D., & Diamond, I. T. The neural basis of auditory discrimination. In H. F. Harlow & C. N. Woolsey (Eds.), *Biological and biochemical bases of behavior.* Madison: Univer. Wisconsin Press, 1958.

Neff, W. D., Fisher, J. F., Diamond, I. T., & Yela, M. Role of auditory cortex in discrimination requiring localization of sound in space. *J. Neurophysiol.*, 1956, **19**, 500–512.

Pavlov, I. P. *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex.* (Trans. by G. V. Anrep) London: Oxford Univer. Press, 1927.

Penfield, W., & Rasmussen, T. *The cerebral cortex of man.* New York: Macmillan, 1950.

Pestalozza, G., & Davis, H. Electric responses of the guinea pig ear to high audio frequencies. *Amer. J. Physiol.*, 1956, **185**, 595–609.

Pierce, A. H. *Studies in auditory and visual space perception.* New York: Longmans, Green, 1901.

Piéron, H. L'orientation auditive latérale. *Année psychol.*, 1922, **23**, 186–214.

Rayleigh, Lord. Acoustical observations. *Phil. Mag.*, 1877, **3**, 456–464.

Rayleigh, Lord. On our perception of sound direction. *Phil. Mag.*, 1907, **13**, 214–232.

Riss, W. Effect of bilateral temporal cortical ablation on discrimination of sound direction. *J. Neurophysiol.*, 1959, **22**, 374–384.

Rosenblith, W. A., & Rosenzweig, M. R. Electrical responses to acoustic clicks: Influence of electrode location in cats. *J. Acoust. Soc. Amer.*, 1951, **23**, 583–588.

Rosenzweig, M. R. Representations of the two ears at the auditory cortex. *Amer. J. Physiol.*, 1951, **167**, 147–158.

Rosenzweig, M. R. Cortical correlates of auditory localization and of related perceptual phenomena. *J. comp. physiol. Psychol.*, 1954, **47**, 269–276.

Rosenzweig, M. R., & Amon, A. H. Binaural interaction in the medulla of the cat. *Experientia*, 1955, **11**, 498–500.

Rosenzweig, M. R., & Rosenblith, W. A. Some electrophysiological correlates of the perception of successive clicks. *J. Acoust. Soc. Amer.*, 1950, **22**, 878–880.

Rosenzweig, M. R., & Sutton, D. Binaural interaction in lateral lemniscus of cat. *J. Neurophysiol.*, 1958, **21**, 17–23.

Rosenzweig, M. R., & Wyers, E. J. Binaural interaction at the inferior colliculi. *J. comp. physiol. Psychol.*, 1955, **48**, 426–431.

Spallanzani, M. Observations on the organs of vision in bats. *Phil. Mag.*, 1798, **1**, 134–136.

Stenger, —. Simulation und Dissimulation von Ohrkrankheiten und deren Feststellung. *Dtsch. med. Wschr.*, 1907, **33**, 970–973.

Stevens, S. S., & Davis, H. *Hearing: Its psychology and physiology.* New York: Wiley, 1938.

Stotler, W. A. An experimental study of the cells and connections of the superior olivary complex of the cat. *J. comp. Neurol.*, 1953, **98**, 401–432.

Supa, M., Cotzin, M., & Dallenbach, K. M. "Facial vision": The perception of obstacles by the blind. *Amer. J. Psychol.*, 1944, **57**, 133–183.

ten Cate, J. Akustische und optische Reaktionen der Katzen nach teilweisen und totalen Extirpationen des Neopallismus. *Arch. Néerl. Physiol.*, 1934, **19**, 191–264.

Teuber, H. L., & Diamond, S. Effects of brain injury on binaural localization of sounds. Paper read at Eastern Psychological Association, Atlantic City, March 1956.

Thompson, S. P. Phenomena of binaural audition. *Phil. Mag.*, 1877, **4**, 274–276.

Thompson, S. P. Phenomena of binaural audition. *Phil. Mag.*, 1878, **6**, 383–391.

Trimble, O. C. The theory of sound localization: A restatement. *Psychol. Rev.*, 1928, **35**, 515–523.

Tunturi, A. R. Audiofrequency localization in the acoustic cortex of the dog. *Amer. J. Physiol.*, 1944, **141**, 397–403.

Tunturi, A. R. A study of the pathway from the medial geniculate body to the acoustic cortex in the dog. *Amer. J. Physiol.*, 1946, **147**, 311–319.

Upton, M. Differential sensitivity in sound localization. *Proc. Nat. Acad. Sci., Wash.*, 1936, **22**, 409–412.

Venturi, G. Riflessioni sulla conoscenza dello spazio che noi passiamo ricavar dall'udito. In, *Indagine fisica sui colori.* Modena: Società Tipografica, 1801.

Venturi, J. B. Considerations sur la connaissance de l'étendue que nous donne le sens de l'ouie. *Mag. encycl.* ou *J. Lett. Arts* (later title: *Ann. encycl.*), 1796, **3**, 29–37.

Venturi, J. B. Betrachtungen über die Erkenntnis der Entfernung, die wir durch das Werkzeug des Gehörs erhalten. *Arch. Physiol.* (*Voigt's Mag.*), 1800, **5**, 383–392. (a)

Venturi, J. B. Betrachtungen über die Erkenntniss des Raums, durch den Sinn des Gehörs. *Mag. neu. Zustand Naturkd.* (*Reil's Arch.*), 1800, **2**, 1–16. (b)

von Békésy, G. Ueber das Richtunghören bei einer Zeitdifferenz oder Lautstarkenungleichheit der beiderseitigen Schalleinwerkungen. *Phys. Z.*, 1930, **31**, 824–835, 857–868.

von Hornbostel, E. M. Das räumliche Hören. In A. Bethe (Ed.), *Handbuch der normalen und pathologischen Physiologie.* Vol. II. Berlin: Springer, 1926.

von Hornbostel, E. M., & Wertheimer, M. Ueber der Wahrnehmung der Schallrichtung. *SB Preuss. Akad. Wiss.*, 1920, 388–396.

Walsh, E. G. An investigation of sound localization in patients with neurological abnormalities. *Brain*, 1957, **80**, 222–250.

Watson, L. A., & Tolan, T. *Hearing tests and hearing instruments.* Baltimore: Williams & Wilkins, 1949.

Woodworth, R. S. *Experimental psychology.* New York: Holt, 1938.

Woolsey, C. N., & Walzl, E. M. Topical projection of nerve fibers from local regions of the cochlea to the cerebral cortex of the cat. *Bull. Johns Hopkins Hosp.*, 1942, **71**, 315–344.

(Received April 18, 1960)

Psychological Bulletin
1961, Vol. 58, No. 5, 390-419

PHYSIOLOGICAL EFFECTS OF "HYPNOSIS"[1]

THEODORE XENOPHON BARBER

Worcester Foundation for Experimental Biology and Medfield State Hospital, Massachusetts

This paper reviews two series of investigations: one series indicating that sensory, circulatory, gastrointestinal, and cutaneous functions can be altered by means of "hypnosis"; and a second series indicating that similar physiological effects can be produced by symbolic stimulation without "hypnosis."

Sensory Alterations Induced by Hypnotic Stimulation[2]

"*Hypnotic Color-Blindness*"

To induce "color-blindness" in six "trained" hypnotic subjects (*S*s), Erickson (1939) employed a complex procedure which included the following: gradual induction of a "profound somnambulistic hypnotic trance"; slow, gradual induction of "total blindness"; awakening of the *S* in the "blind" condition so that he would experience distress and anxiety; induction of a second "trance" condition; explanations to the *S* that vision would be restored but that a certain color or colors would not be detectable; suggestions of amnesia for the critical color or colors; administration of the Ishihara during suggested (green, red, red-green, and total) color-blindness; administration of the Ishihara without suggested color-blindness in the waking state and in "the simple trance state." The results of this involved experiment (which included 13 separate administrations of the Ishihara to each *S*) appeared to be as follows: all *S*s had normal color vision during the waking state and in "the simple trance state"; during suggested color-blindness, the numerals on some of the Ishihara cards were read in the manner characteristic of the green, red, red-green, or total color-blind. Erickson concluded that the hypnotic procedure was effective in inducing "consistent deficiencies in color vision comparable in degree and character with those found in actual color blindness." However, Grether (1940) criticized this conclusion noting that (*a*) "red-green color-blindness" does not exist in nature (this is a generic term referring to symptoms common to red-blindness and green-blindness); and (*b*) the deficiencies in color vision found among persons with actual red-blindness, green-blindness, or total color-blindness are "quite different" from those which Erickson attempted to induce. Harriman (1942) repeated part of Erickson's procedure, suggesting amnesia for red and then for green to 10 "deeply hypnotized" *S*s; although these *S*s responded to the Ishihara in a manner similar to Erickson's *S*s, Harriman concluded, in accordance

[1] The author is indebted to Louis B. Glass, Alejandro D. Paniagua, David R. Evans, and Harry Freeman for critically reading the manuscript. A condensed version of the paper was presented at the University of Kansas seventh Annual Institute of Research in Clinical Psychology, May 10, 1960.

Research reported by the author and his associates was supported in part by a grant from the National Association for Mental Health and in part by Research Grant MY3253, from the National Institute of Mental Health, United States Public Health Service.

[2] Since experimental and clinical studies of "hypnotic analgesia" have been recently reviewed elsewhere (Barber, 1959, 1960a), this phenomenon is not included in the following discussion.

with Grether's critique, that the alterations induced "resemble attitudinal changes more closely than they resemble profound changes in sensory content." However, no attempt was made to determine if the lengthy and involved hypnotic procedure employed in the investigation was actually necessary to induce such "visual anomalies."

Barber and Deeley (1961) hypothesized that normal *S*s, instructed to remain inattentive to red or green, give similar responses to the Ishihara as "hypnotic color-blind" *S*s. As a preliminary test of color vision, the American Optical Company Pseudo-Isochromatic Plates were administered to 10 normal *S*s. The *S* was then presented with the Ishihara plates and instructed as follows: "Now look at these cards. As I present each card, try as hard as you possibly can to pay no attention to the red. Look carefully at the rest of the card, but ignore the red; just don't let yourself see it." After completing this task the Ishihara cards were presented again and similar instructions were given to "try as hard as you possibly can to pay no attention to the green." Finally, the *S* was instructed to report what he naturally saw on the Ishihara plates. The results were as follows: (*a*) The responses to the Pseudo-Isochromatic Plates and to the final administration of the Ishihara indicated normal color vision in all *S*s. (*b*) When instructed to "pay no attention" to red and then to green, 92 of 320 (28.8%) of the total responses of the 10 normal *S*s were similar to the responses expected from persons with natural red-blindness or green-blindness. Of the 320 responses given to the Ishihara by Harriman's 10 "deeply hypnotized" *S*s during suggested red-blindness and green-blindness, 85 (or 26.6%) were similar to the responses expected from the red-blind or green-blind. In brief, this experiment appears to indicate that normal persons who have been instructed to concentrate away from red or green give similar responses on the Ishihara as "deeply hypnotized" *S*s who have been given elaborate suggestions to induce color-blindness.

"*Hypnotic Blindness*"

Are hypnotic suggestions of total blindness effective in altering physiological processes related to vision? Hernàndez-Peòn and Donoso (1959) recently published a neurophysiological experiment which, although not a direct study of hypnotically-induced blindness, nevertheless promises to contribute to our understanding of this phenomenon. Electrodes were deeply implanted in the occipital lobes of five patients who had undergone trephination for diagnostic explorations. With the occipital electrodes in place, the room was darkened and the patient was stimulated by electronic lamp flashes at the rate of 1/millisecond. In each case the electrographic recordings showed an evoked potential simultaneous with the photic stimulation. Subsequently, when two of the patients, whom the experimenters judged to be especially "suggestible," were given repeated verbal suggestions that the light intensity was greater than that actually applied, the electrographic recordings indicated an enhancement of the photically evoked potentials; when given the suggestion that the intensity of the light had diminished, while it actually remained constant, the recordings showed a diminution of the evoked potentials. However, in related experiments the same investigators demonstrated that the magnitude of the photically evoked potentials was consistently reduced whenever "the attention of the subject was distracted," e.g., when instructed to solve a difficult arithmetic problem mentally or when asked to recall an interesting experience. From these experiments and from a series of related studies by other workers summarized in the paper, the authors suggest that during

> "voluntary attention" as well as by suggestion, transmission of photic impulses is modified at the retina by centrifugal influences. These influences, acting during wakefulness, are probably related to organized activity of the reticular formation of the brain stem under the control of the cortex (p. 394).

In earlier studies, Dorcus (1937), Lundholm and Lowenbach (1942), and other workers had noted that the pupillary reaction to light stimulation is not altered during "hypnotic blindness." However, since pupillary constriction to light is found during some types of organic blindness (e.g., bilateral destruction of the occipital visual areas—Madow, 1958), this response is not a satisfactory index of blindness and workers in this area have generally focused on an ostensibly more satisfactory response—alpha blocking on the electroencephalogram (EEG).

Alpha blocking to photic stimulation appears to be a totally involuntary response which is almost always present in normal persons and never present in the blind. A series of investigations has demonstrated that (*a*) when the room is darkened and the eyes are closed, most normal persons typically show an alpha rhythm on the EEG (consisting of waves with a frequency of 8 to 13 cycles per second and an amplitude of about 50 microvolts); (*b*) a light flashed into the closed eyes of these individuals is almost always effective in causing "alpha block" or "alpha desynchronization" (i.e., in replacing the alpha rhythm with small fast waves) within 0.4 second (Jasper & Carmichael, 1935); and (*c*) persons with total blindness of neurological origin do not show alpha blocking under these conditions (Callahan & Redlich, 1946).

Lundholm and Lowenbach (1942), Barker and Burgwin (1948), and Ford and Yeager (1948) found that hypnotic suggestions of blindness did *not* prevent alpha blocking when the *S*s opened their eyes in an illuminated room. However, these experiments are based on a methodological error: In normal persons the act of opening the eyes per se—whether in darkness or in an illuminated room—almost invariably results in alpha desynchronization (Loomis, Harvey, & Hobart, 1936; Yeager & Larsen, 1957). To determine if hypnotic suggestions of blindness are effective in preventing the alpha desynchronization which normally occurs after visual stimulation, it is therefore necessary for the *S* either to keep his eyes continuously open or continuously closed during the experiment. These conditions have been met in three investigations. Loomis et al. (1936) demonstrated that when total blindness was suggested to an excellent hypnotic *S* whose eyes were kept open continuously with adhesive tape, the alpha rhythm did *not* show desynchronization during photic stimulation. This was repeated 16 times with the same results; whether the room was illuminated or darkened made no difference whatsoever—the alpha rhythm was continuously present until the *S* was told that he could once again see. In a subsequent experiment, Schwarz, Bickford, and Rasmussen (1955) found that after suggestions of blindness 7 of 11 hypnotic *S*s (with eyes taped open) showed occasional alpha waves when the room was illuminated. In a more recent study, Yeager and Larsen (1957) instructed five *S*s to keep their eyes continuously closed during the experiment. Hypnotic and posthypnotic suggestions were given that the *S* would not be aware of the light stimulation. In the majority of trials, no alpha blocking occurred when light fell upon the closed eyes.

The above studies indicate that hypnotic suggestions of blindness are at times effective in eliminating an involuntary physiological response which normally follows visual stimulation, viz., alpha blocking on the electroencephalogram. However, a similar effect can be demonstrated in *S*s who have not been given an "hypnotic induction" and who do not appear to be in "the trance state." Loomis et al. (1936) found that when a uniformly illuminated bowl was placed over the eyes of a normal person who was instructed not to focus on any specific part of the light pattern, the alpha waves appeared fairly regularly. Gerard (1951) writes:

> With a little practice I can look directly at a 100-watt light . . . and, by deliberately paying no attention to it, I can have my alpha waves remain perfectly intact; then with no change except what I can describe in no other way than as directing my attention to the light, have them immediately disappear (p. 94). (Quoted by permission of John Wiley & Sons.)

Jasper and Cruikshank (1937) have published similar findings. In brief, although some "hypnotized" *S*s, who have been given suggestions of blindness, continue to show an occipital alpha rhythm part of the time or all of the time when stimulated by light, a similar effect can be demonstrated in normal persons who are instructed to "pay no attention" to visual stimuli.

In a recent study Schwarz et al. (1955) found that five "hypnotized" *S*s who had been given suggestions of blindness did not show eye movements when urged to look at an object. The restriction of eye movements was indicated both by electromyographic eye leads and by the marked suppression of lambda waves on the EEG. These investigators suggest that the restriction of eye movements during hypnotic blindness "is an attempt to shut off all alerting stimuli that might interfere with the successful accomplishment of thesuggestion." Along similar lines, Barber (1958b) presented evidence indicating that seven somnambulistic hypnotic *S*s deliberately refused to look at an object which they had been told that they could not see; observation of eye movements indicated that they typically focused on all parts of the room *except* where the object was situated. When interviewed after the experiment, most of the *S*s readily admitted that they purposely refused to carry out the active process of turning the head and focusing the eyes on the object, e.g., "I was almost carefully not looking at it," "I kept looking around it or not on it."

In an earlier study, Pattie (1935) gave five good hypnotic *S*s the suggestion that they were blind in one eye. Four responded to a series of visual tests (stereoscope, perimetry, filters, Flees' box, plotting the blind spot, opthalmological examination) with normal vision in both eyes; however, one *S* responded to all tests as if she were actually blind in one eye. In a second experiment the "blind" *S* was given a more complicated filter test; the results indicated that the "blind" eye was not impaired to the slightest degree and Pattie concluded that the "former tests were thus invalidated." When questioned in a subsequent hypnotic session, the *S* revealed, after much resistance, that she had given a convincing demonstration of uniocular blindness because of the following: during the stereoscopic test the two images were separated a second after exposure and this gave her the necessary knowledge to fake the test; she had practiced determining the blind spot at home after the experimenter had first attempted to plot it; on the Flees' box with crossed images she "saw there were mirrors in there and figured somehow that the one on the left was supposed to be seen with the right eye," etc.

The above studies appear to indicate that the "good" hypnotic *S*, who has been given suggestions of blindness, purposely attempts to inhibit responses to visual stimuli. This suggests the following hypothesis which can be easily confirmed or disproved: The responses to photic stimulation which characterize "deeply hypnotized" *S*s who have been given suggestions of blindness can be duplicated by normal persons who are asked to remain inattentive and unresponsive to visual stimuli.

"Hypnotic Deafness"

Can significant alterations in auditory functions be demonstrated in the hypnotized person following suggestions of deafness? Fisher (1932) and Erickson (1938b) approached this question by investigating the effect of hypnotically-induced "deafness" on conditioned responses to acoustic stimuli. Fisher found that during posthypnotic deafness one *S* did not show a patellar response which had been conditioned to the sound of a bell; Erickson similarly demonstrated that after hypnotic suggestions of deafness two *S*s failed to show a hand-withdrawal response conditioned to the sound of a buzzer. Although both investigators interpret the failure to show conditioned responses to auditory stimuli as a sign of deafness, earlier experiments, reviewed by Hilgard and Marquis (1940, p. 35, pp. 269–270), which indicate that such conditioned responses can be voluntarily inhibited, suggest a second interpretation, namely, that the "hypnotic deaf" *S*s perceived the sound stimulus but purposely inhibited the response. Some support for this interpretation is offered by the kymographic tracings reproduced in Fisher's paper which show an aborted patellar response to some of the sound stimuli. Additional evidence is presented by Lundholm (1928) who, like Erickson, conditioned a hand-withdrawal response to an auditory stimulus; although the *S* in this case did not show the conditioned response after hypnotic suggestions of deafness, he later admitted "having heard the click, having felt an impulse to withdraw on click without shock, and having resisted and inhibited that impulse" (p. 340).

As an additional index of deafness, Erickson (1938a) noted that his *S*s did not show startle responses to sudden loud sounds. Other investigations, however, again suggest the possibility that the *S*s may have perceived the sound and purposely inhibited the startle response; for instance, Dynes (1932) reported that three "hypnotic deaf" *S*s, who did not show overt startle responses when a pistol was unexpectedly fired, admitted after the experiment that they heard the sound, and Kline, Guze, and Haggerty (1954) demonstrated that a "hypnotic deaf" *S* who failed to show both conditioned responses to auditory stimuli and startle reflexes to sudden loud sounds showed clear-cut responses to auditory stimuli when tested by a method employing delayed speech feedback.

The latter experiment merits further comment. In the normal person, feeding back his speech through tape recording amplification and earphones with a delay of one-quarter second has been reported to produce an impairment in subsequent speech. Most commonly this speech disturbance involves stammering, stuttering, perseveration, and marked loss in speed and tempo. Kline et al. (1954) found that such delayed speech feedback produced distinct impairment in speech performance in an excellent hypnotic *S* who had been given suggestions of deafness. However, as compared with his "waking" performance, the *S* showed less slurring, stuttering, and stammering, appeared more calm, and did not show discomfort. The investigators concluded that the hypnotic suggestions of deafness were effective in inducing a "set," or in "gearing" the *S*, "to give minimal response to the excruciating intensity and the constant interference of the feed-back of his own voice" without in any way inducing "deafness in the usual sense." However, no attempt was made to determine if the *S* would have shown a similar ability to tolerate the speech-disturbing stimulation during the "waking" experiment if he had been carefully instructed and motivated to remain inattentive to or to "concentrate away from" the stimulation. Further experiments are required to determine if normal persons are able to duplicate the behavior of this "deeply hypnotized" *S* when instructed in this manner.

Malmo, Boag, and Raginsky (1954) have reported comparable findings. After appropriate suggestions to induce deafness, two somnambulistic *S*s denied auditory sensations and showed significantly reduced motor reactions to sudden auditory stimulation; however, myographic recordings from eye muscles showed a strong blink reaction in both *S*s at each presentation of the auditory stimulus. Sternomastoid tracings indicated that one *S* showed slight startle responses to all stimuli and the other *S* showed a strong startle reaction to the first presentation of the stimulus and slight startle reactions to subsequent stimuli. Other data presented in the report (e.g., introspective reports and myographic tracings indicating a higher level of tension in the chin muscles under hypnosis as compared to the control condition) permit the following interpretation of the findings: (*a*) the *S*s were unable to inhibit blink responses to the auditory stimuli; (*b*) since the first presentation of the auditory stimulus was more or less unexpected, one *S* failed to inhibit the startle response; (*c*) since the second and subsequent stimuli were expected, both *S*s were able, to a great extent, to inhibit startle responses. In an earlier study Malmo and his collaborators (Malmo, Davis, & Barza, 1952) found that, when unexpectedly presented with an intense auditory stimulus, a hysterical "deaf" patient also showed a gross startle response on the myograph; a control case of middle-ear deafness, studied by the same techniques, showed no blink reaction and no startle response to any presentation of the auditory stimulus.

In an earlier study Pattie (1950) gave four somnambulistic hypnotic *S*s suggestions of unilateral deafness.

The *S*s appeared to accept the suggestions, insisting that they could not hear in one ear. However, when auditory stimuli were presented in such a manner that they could not determine which ear was being stimulated, they showed normal hearing in both ears.

The above findings—that "hypnotic deaf" *S*s purposely inhibit conditioned responses to auditory stimuli (Lundholm), appear to inhibit startle responses to sudden acoustic stimuli (Malmo et al., 1952), show a calmer attitude and less tension during speech-disturbing auditory stimulation but no sign of actual deafness (Kline et al., 1954), and do not show "deafness" in one ear when unable to determine which ear is being stimulated (Pattie, 1950)—suggest a similar hypothesis as the studies of "hypnotic blindness" reviewed in the preceding section of this paper: if carefully instructed and motivated to "concentrate away from" auditory stimulation, normal persons show similar responses to acoustic stimuli as "hypnotic deaf" *S*s.

The Effect of Hypnotic Stimulation on Circulatory Functions

Effect of Hypnotic Stimulation on Vasomotor Functions

The evidence at present indicates that localized vasoconstriction and vasodilation (and a concomitant localized skin temperature alteration) can be induced in some hypnotized persons by appropriate verbal stimulation. McDowell (1959) found that a good hypnotic *S* showed erythema with vasodilation and increase in skin temperature of the right leg following suggestions that the leg was immersed in warm water. In a careful experiment, Chapman, Goodell, and Wolff (1959) suggested to 13 *S*s "as soon as a state of moderate to deep hypnosis had been established," that one arm was either "normal" or that it was numb, wooden, and devoid of sensation ("anesthetic"). The arm was then exposed on three spots, blackened with India ink, to a standard thermal stimulus (500 millicalories/second/centimeter^2 for 3 seconds). After an interval of 15 to 30 minutes "during which [time] hypnosis was continued," it was suggested that the other arm was tender, painful, burning, damaged, and exceedingly sensitive ("vulnerable") and the same standard noxious stimulation was applied. The results of 40 experiments with the 13 *S*s were as follows: In 30 experiments the inflammatory reaction and tissue damage following the noxious stimulation was greater in the "vulnerable" arm, in 2 experiments the reaction was greater in the "anesthetic" arm, and in 8 experiments no difference was noted. Plethysmographic and skin temperature recordings indicated that following the noxious stimlation local vasodilation and elevation in skin temperature was larger in magnitude and persisted longer in the "vulnerable" arm. This experiment should be repeated with unhypnotized *S*s who are instructed to *imagine* one arm as "devoid of sensation" and the other arm as "exceedingly sensitive." The data summarized below suggest that at least some of the effects reported in this study—localized vasodilation and elevation in skin temperature—can be induced by symbolic stimulation in some individuals who do not appear to be "in a state of moderate to deep hypnosis."

When attempting to condition local vasoconstriction and vasodilation to verbal stimuli, Menzies (1941) found that the conditioning procedure could be dispensed with in some cases; some persons, who had *not* participated in the experimental conditioning, showed vasodilation in a limb when recalling previous experiences involving warmth of the limb and local vasoconstriction when recalling experiences involving cold. In an earlier study, Hadfield (1920) found that localized changes in skin temperature could be induced by suggestions given to a person "in the waking state." In this case, the *S* had exercised vigorously before the experiment and the temperature of both hands, as measured with the bulb of the thermometer held firmly in the palm, had reached 95°F. Without a preliminary hypnotic procedure, it was suggested that the right arm was becoming cold. Within half an hour the temperature of the right palm fell to 68° while the temperature of the left palm remained at 94°. When subsequently given the suggestion that the right hand was becoming warm, the temperature of the hand rose within 20 minutes to 94°. Although this *S* had previously participated in hypnotic experiments, Hadfield insists that he did not "hypnotize" him during this experiment and that the temperature alterations occurred when the *S* was "entirely in the waking condition."

Cardiac Acceleration Produced by Hypnotic Stimulation

A number of experiments, reviewed by Gorton (1949) and Weitzenhoffer (1953), demonstrate that cardiac acceleration can be produced by hypnotic suggestions which activate the *S* and that cardiac deceleration can be produced by hypnotic suggestions of relaxation, drowsiness, and sleep; however, this finding indicates no more and no less than that an alteration in the "level of arousal" or "level of activation" (Duffy, 1957; Woodworth & Schlosberg, 1954)—whether induced by stimuli present during various ongoing life situations or induced by symbolic stimulation during a hypnotic experiment—is correlated with an alteration in the heart rate. A more significant question is: Can the heart rate be accelerated or depressed by direct suggestions of such an effect without simultaneously inducing anxiety, emotion, or arousal? Solovey and Milechnin (1957) demonstrated an accelerated pulse rate in 2 out of 23 hypnotic *S*s following the direct suggestion "Your heart is beating more rapidly." However, the possibility is not excluded that the cardiac acceleration in the two cases was due to emotion or anxiety evoked by the suggestions; one *S* later reported that, when given the suggestion, he imagined himself looking down from a height and feeling someone pushing him on the shoulder and the other *S* stated that, when given the suggestion, he had a feeling of distress. Since relatively large changes in cardiac rate can be demonstrated during alterations in the rate and depth of respiration (Huttenlocher & Westcott, 1957), it also appears plausible that the altered pulse rate in these cases may have been an indirect effect of a change in respiratory pattern.

To demonstrate a *direct* effect of symbolic stimulation on heart rate it is necessary to control at least two factors, "level of arousal" and respiratory rate. To the writer's knowledge only one hypnotic experiment has been published which ostensibly satisfies these criteria: Van Pelt (1954) reported that a somnambulistic hypnotic *S* showed an accelerated cardiac rate following direct suggestions of such an effect without at the same time showing an altered respiratory rate or emotional arousal. After an "hypnotic induction" procedure, this investigator spoke to the *S* in a quiet tone as follows: "Your heart is beginning to beat faster. It is getting faster and faster. You are perfectly calm, but your heart is beating faster and faster." The electrocardiogram (EKG) showed that the heart rate increased immediately from 78 to 135 beats per minute. Although Van Pelt states that he did not observe a change in the depth and rate of respiration during the tachycardia, it appears possible that an altered respiratory pattern could have been demonstrated if a pneumograph had been employed in the study. However, during the acceleration the *S* appeared calm and the EKG tracing did not show somatic tremors which are typical of nervousness and fear. In a second experiment, in which the same *S* showed cardiac acceleration following suggestions intended to arouse fear—"You are driving a car at a tremendous speed and are heading toward a second car and are going to crash"—the EKG recording showed clear evidence of somatic tremors.

The above study lacks a crucial control; no attempt was made to determine if the *S* could voluntarily accelerate the heart without "hypnosis." Since other workers employing similar procedures with equally "good" hypnotic *S*s have failed to demonstrate cardiac acceleration (e.g., Jenness & Wible, 1937, failed in 30 attempts with eight somnambulists), it appears plausible that the hypnotic procedure was not a necessary factor in producing this effect. Supporting evidence for this supposition is presented in a series of studies (ca. 20) which demonstrate that some apparently normal persons are able to accelerate the heart voluntarily (King, 1920). In most of the reported cases the voluntary tachycardia was on the order of 15 to 40 beats per minute; however, in some cases (Favell & White, 1917; Tarchanoff, 1885) the acceleration was as high as 63 or 75 beats per minute. In all cases the *S*s denied that they induced the tachycardia by visualizing emotion inducing situations and insisted that they produced the effect by voluntary effort. Some *S*s showed changes in respiratory pattern during the voluntary tachycardia but in these cases the respiratory alterations varied and could not be correlated with the change in heart rate (Koehler, 1914; Pease, 1889; Tarchanoff, 1885; Van de Velde, 1897); other *S*s could as readily induce the voluntary acceleration when breathing more or less normally as when showing changes in respiratory pattern (King, 1920; Taylor & Cameron, 1922); and some *S*s showed no significant change in respiratory pattern on the pneumograph when inducing cardiac acceleration on the order of 40 beats per minute (Favill & White, 1917).

Voluntary acceleration of the heart may not be as uncommon as is generally assumed: Van de Velde found four cases and Tarchanoff five cases when confining their search to relatively small groups of individuals; a number of medical students discovered that they possessed this ability in physiology classes when they attempted to determine the validity of the lecturer's assertion that voluntary cardiac acceleration is not impossible (Ogden & Shock, 1939; West & Savage, 1918).

Cardiac Standstill Induced by "Hypnosis"

Raginsky (1959) demonstrated that hypnotic suggestions are effective in producing cardiac block for a brief period in an appropriately predisposed person. The *S* in this case was a hospitalized patient whose carotid sinuses had been surgically removed because of periodic fainting episodes with cardiac arrest (Adams-Stokes disease). After the patient "went into a medium to deep hypnotic state," he was instructed "in a tone of considerable urgency to visualize with all clarity possible his worst attack of faintness." The patient "turned pale, limp, and a cold perspiration appeared on his forehead. His pulse was unobtainable . . . ." The EKG tracing showed complete auricular and ventricular standstill for a time interval of four beats, followed by a normal sinu-auricular beat. After a rest period of 10 minutes, the experiment was repeated with comparable results. However, no attempt was made to determine if cardiac arrest could be induced in this patient by asking him to visualize his worst attack of faintness *without* a preceding "hypnotic induction" procedure. The case summarized below suggests that the "hypnotic induction" and the "medium to deep hypnotic state" may have been unnecessary in producing this effect.

McClure (1959) found that an appropriately predisposed person could voluntarily produce cardiac standstill. The *S* in this case, a 44-year-old airplane mechanic, had discovered that he could induce a progressive slowing of the pulse by relaxing completely. When asked to induce a diminution of the heart rate in the laboratory, the *S* lay very quitely and allowed respiration to become extremely shallow. The EKG showed sinus arrest for a period of 5 seconds. An EKG tracing taken 1 hour after the experiment was within normal limits. Since the *S* had rheumatic fever at age 7, McClure suggests the following tentative explanation of this performance:

> The underlying cardiac change is believed to be well compensated rheumatic heart disease with aortic valvulitis. The bradycardia and cardiac arrest are probably manifestations of exaggerated vagotonia, induced through some mechanism which, although under voluntary control, is not known to the patient himself. Careful observation did not reveal any breath-holding or Valsalva maneuver in connection with the cessation of heartbeat. Apparently the patient simply abolished all sympathetic tone by complete mental and physical relaxation (pp. 440–441). (Quoted by permission of *California Medicine*.)

Electrocardiographic Alterations Induced by Hypnotic Stimulation

Bennett and Scott (1949) found that one of five excellent hypnotic *S*s showed tachycardia and T wave abnormalities on the EKG—lowering or disappearance of T in Leads I, II, and III—within 2 minutes following suggestions intended to induce anxiety and anger. The *S* in this case was an emotionally stable and well-adjusted young male with no history of cardiac disorders and with an otherwise normal EKG. Since such EKG abnormalities are not normally associated with tachycardia, two electrocardiographers, who were not informed of the experimental conditions under which the tracings were made, interpreted the records as indicating coronary artery disease or acute rheumatic fever. Finding in a subsequent study with the same *S* that subcutaneous administration of epinephrine elicited lower T waves in Leads I and II than those found during the control experiment, the authors suggest that the EKG alterations induced during the hypnotic experiment may have been an indirect result of sympathetic stimulation and release of epinephrine from the adrenal medulla. Berman, Simonson, and Heron (1954) confirmed this study; employing 14 susceptible hypnotic *S*s with otherwise normal EKG, they found that during hypnotically-induced fear and anxiety, two showed elevation and five showed depression or inversion of T waves. In a second experiment these workers found that although "deep hypnosis" could not be induced in 11 patients with coronary sclerosis and angina pectoris, four showed T wave changes when given emotion inducing suggestions.

The above experiments demonstrate that EKG alterations resembling those found in grave cardiac disorders can be induced in some "hypnotized" *S*s by suggestions which evoke fear, anger, or anxiety; however, similar EKG abnormalities have been demonstrated in some normal persons during emotional arousal. Mainzer and Krause (1940) compared the EKG tracings of 53 unselected surgical patients recorded the day before surgery, and on the operating table immediately before the induction of general anesthesia. As compared with the earlier tracings, 40% of the tracings recorded immediately before surgery showed various abnormalities such as S-T depression with T low, inverted, or absent. Along similar lines, Landis and Slight (1929) and Loftus, Gold, and Diethelm (1945) demonstrated that some persons with otherwise normal EKG show abnormalities of the ST segment and the T wave during startle or anxiety; Crede, Chivers, and Shapiro (1951) found that in rare cases mere anticipation of the EKG test is sufficient to produce inverted T waves in normal individuals; and Ljung (1949) published a study of 14 *S*s with no evidence of cardiac disease who showed abnormal T waves during apparently slight emotional stimulation. After summarizing these and related investigations, Weiss (1956) suggests that such EKG effects are found during emotional stimulation in persons who are prone to show an elevation of sympathetic tone and an increase in cardiac metabolism without a corresponding increase in the coronary circulation.

In brief, the above studies on cardiac functions indicate the following:

1. In very rare cases, it is possible to produce cardiac acceleration or complete stoppage of the heart action by appropriate hypnotic stimulation; however, in very rare cases, similar effects can be voluntarily produced by unhypnotized persons.

2. Although some hypnotized persons show EKG alterations resembling those found in organic heart disease following suggestions designed to induce fear, anxiety, or anger, some persons who have not been given an "hypnotic induction" and who do not appear to be "in trance" show similar EKG alterations during emotional stimulation.

Effect of Hypnotic Stimulation on Metabolic and Gastrointestinal Functions

Effect of Hypnotic Stimulation on Blood Glucose Levels

A number of experiments appear to indicate that hypnotized persons show an elevation of blood glucose levels when given the direct suggestion that blood sugar will increase. Before discussing these studies, it is appropriate to note the following:

1. The level of blood glucose appears to be closely related to the level of "arousal"; blood sugar tends to increase during anxiety, emotion, or maintained activity and to decrease during relaxation, depression, or sleep (Dunbar, 1954, Ch. 8).

2. The blood glucose level is excessively labile in diabetics, i.e., as

compared with normal persons, diabetics tend to show more extreme alterations in blood sugar content during periods of high or low "arousal" (Hinkle & Wolf, 1953; Mirsky, 1948).

The above postulates suggest that in diabetic patients any procedure (hypnotic or nonhypnotic) which induces relaxation or minimizes excitability will tend to depress the blood sugar level and minimize glycosuria and any procedure which induces arousal or excitability will tend to elevate the blood glucose level and increase glycosuria. Data supporting this hypothesis have been presented by Bauch (1935) in a study of the effects of training in relaxation (Schultz's "autogene training") on seven diabetic patients. Each patient showed a significant decrease in blood sugar levels after becoming proficient in inducing relaxation—insulin dosage was reduced in each case by 10 to 20 units. Apparently, healthy persons do *not* show the same degree of reduction in blood glucose content after achieving the same success in producing relaxation (Schultz & Luthe, 1959). Along similar lines, Mohr (1925) relieved a "full-pledged diabetic" of glycosuria by hypnotic suggestions which were effective in mitigating his "affective excitability" toward certain significant persons in his surroundings and was able to reinstate the glycosuria by suggesting that he would again be upset by these people. This experiment was repeated four times with the same results.

With the above findings in mind, the results reported in two hypnotic experiments become less mysterious. Gigon, Aigner, and Brauch (1926) found that blood sugar tended to be reduced in four hypnotized diabetic patients following suggestions of relaxation and suggestions that "the pancreas would secrete insulin and that blood and urine sugar would markedly decrease." Although the reduction in blood glucose in these cases may have been due to the suggestion that "the pancreas would secrete insulin," it appears equally plausible that it was a secondary effect of the suggestions of relaxation. Along similar lines, Stein (quoted by Dunbar, 1954, p. 291) reported that direct suggestions that blood sugar would decrease given to six hypnotized diabetic patients resulted in reduced blood sugar in 47 out of 56 attempts. Again, it appears plausible that the reduced blood glucose in these cases was an indirect result of the suggestions of relaxation given during the "hypnotic induction" procedure. Supporting evidence for this supposition is presented in a second experiment by the same investigator; although Stein used only one diabetic patient in this study, he found that an "hypnotic induction" (apparently consisting of suggestions of quietude, relaxation, and drowsiness) resulted in a significant fall in blood glucose content *without* suggesting that the blood sugar would fall.

Is it possible to elevate the blood sugar level by suggesting to a nondiabetic hypnotic *S* that he is ingesting sugar? Marcus and Sahlgren (1925) found no rise in blood glucose content when four "deeply hypnotized" nondiabetics were given a saccharin solution which they were told was a sugar solution. Similarly, Nielsen and Geert-Jorgensen (1928) found no elevation in the fasting blood sugar level when six excellent hypnotic *S*s (nondiabetics) were given the suggestion that a glass of water contained large amounts of sugar. In contradistinction to the above, Povorinskij and Finne (1930) found an elevated blood sugar content in two somnambulistic hypnotic *S*s after inducing an hallucination of ingesting sugar and honey; however, an elevation in blood glucose could be demonstrated in one of these *S*s following similar suggestions given during "the waking state." The data presented in the report do not exclude the possibility that the hypnotic suggestions which induced an "hallucination" of ingesting sugar and honey served to "arouse" the subjects or to induce emotional excitement.

Effect of Hypnotic Stimulation on Gastric Functions

The evidence indicates that stomach secretions, hunger contractions, and various other gastrointestinal functions can be influenced by appropriate suggestions given to a hypnotized person. Ikemi (1959) demonstrated that suggestions given during hypnosis of eating a delicious meal resulted in an increase in free acid, total acidity, and quantity of gastric secretions in 34 out of 36 healthy young persons. In an earlier experiment, Heyer (1925) introduced a tube into the stomach of a "deeply hypnotized" *S* and removed the contents. If no secretion occurred within 10 minutes, the *S* was given the suggestion that he was ingesting either meat broth, bread, or milk and the gastric secretions were collected at 5-minute intervals and examined for quantity, acidity, and proteolytic activity. Each of the suggested meals evoked a secretion of approximately 6 to 10 cubic centimeters of "gastric juice" within 10 to 15 minutes and the acidity and proteolytic activity appeared to vary with each food suggested. Delhougne and Hansen (1927) reported a similar study with one somnambulistic *S*. After the *S* was placed in "deep hypnosis," the stomach and duodenal secretions were aspirated by means of a Rehfuss tube. Following this, the *S* was given the suggestion that he was ingesting a meal which was rich in protein (Schnitzel), rich in fat (a biscuit thickly covered with butter), or rich in carbohydrate (chocolate and marchpane). Each of the suggested meals evoked secretions of acid, pepsin, and lipase from the stomach and of trypsin, lipase, and diastase from the pancreas. Although the authors do not analyze the data statistically, they conclude that the hallucinated meals were as effective as actual meals in eliciting *specific* secretions from the stomach and pancreas, e.g., the hallucinated protein meal supposedly induced a specific increase in the secretion of pepsin and trypsin, the hallucinated fatty meal supposedly induced a specific increase in the secretion of lipase. This startling conclusion, however, appears to be erroneous; a statistical analysis indicates that the quantity of each of the enzymes found after the three hallucinated meals was not significantly different.

The above studies do not answer a crucial question: Was the "hypnotic induction" and the appearance of "deep trance" on the part of the *S*s necessary to produce these effects? If the *S*s had been asked to vividly imagine or to think about eating certain foods (without an "hypnotic induction") would they have shown similar pancreatic and gastric secretory activity? That such may have been the case is suggested by an earlier experiment reported by Luckhardt and Johnston (1924). These investigators also found that when a hypnotized *S* was given suggestions of eating a fictitious meal, he showed an increase in the volume and acidity of the digestive secretions comparable to that found when actually eating a meal; however, in the control experiment, when the investigators merely talked to the *S* about an appetizing meal, he showed similar gastric secretory activity. This finding is not unusual. Miller, Bergeim, Rehfuss, and Hawk (1920) reported that the sound and thought alone of a frying steak gave rise to gastric secretory activity in some normal *S*s. Employing a *S* with a gastric fistula, Wolf and Wolff (1947) demonstrated that during the "mere discussion" of eating a certain food the output of hydrochloric acid from the parietal cells was essentially the same as when actually ingesting this food. Similar effects have been demonstrated in other parts of the gastrointestinal tract. Bykov (1957) found that in patients with a gall bladder fistula (but otherwise physiologically normal) "the sight of and even the mere mention of food evoked contraction of the gall bladder" (p. 119). The same investigator also studied a patient with a fistula of the pancreatic duct but otherwise healthy and with a normal digestive tract; 1 or 2 minutes after being drawn into conversation about savory foods, this patient (who was kept on a special diet which served to inhibit secretions) "showed against this inhibitory background abundant pancreatic secretions." (The above patients had *not* participated in experimental conditioning procedures.)

Scantlebury and Patterson (1940) demonstrated that suggestions of eating a fictitious meal were effective in inducing a temporary and at times a complete cessation of gastric hunger contractions in a hypnotized *S*. Lewis and Sarbin (1943) repeated this experiment, employing the Carlson balloon-manometer method with eight *S*s who had fasted prior to the experiment. The *S*s were first given the Friedlander-Sarbin hypnotic induction procedure and rated on "depth of hypnosis." Whenever the *S*s showed gastric hunger contractions, they were given the suggestion of eating a meal. The kymographic tracings showed that the suggestions were effective in inhibiting the hunger contractions in the majority of trials with the "deeply hypnotized" *S*s, in some of the trials with the "moderately hypnotized" *S*s, and in none of the trials with *S*s who were "slightly hypnotized" or not hypnotized. However, a comparable inhibition of hunger contractions could be demonstrated in the "deeply hypnotized" *S*s by asking them to solve an arithmetic problem silently. No attempt was made to determine if hunger contractions could be inhibited in unhypnotized persons by asking them to "vividly imagine" eating a delicious meal.

Earlier studies which did not employ hypnotic procedures found comparable effects. For example, Carlson (1916, p. 152) found that after 4 days of fasting the sight and smell of food inhibited his hunger contractions. Since acid in contact with the gastric mucosa apparently acts reflexly to produce inhibition of gastric contractions (Carlson, 1916, pp. 175–176) and since the "mere thought" of appetizing food gives rise to a significant amount of hydrochloric acid secretion in some normal persons (Miller et al., 1920), it can be hypothesized that suggestions of eating a meal are effective in some "hypnotized" *S*s and some unhypnotized *S*s in inducing gastric acid secretions which act reflexly to inhibit the hunger movements.

In summary, the above studies on metabolic and gastrointestinal functions appear to indicate that blood sugar levels can be altered in diabetic patients by hypnotic or nonhypnotic procedures which alter the level of "arousal," and gastric and pancreatic secretions and gastric hunger contractions can be influenced by symbolic stimulation in both hypnotized and unhypnotized persons.

Effect of Hypnotic Stimulation on Cutaneous Functions

Production of Herpetic Blisters (Cold Sores) by Hypnotic Stimulation

Ullman (1947) reported that a patient (who had been previously cured of hysterical blindness) showed multiple herpetic blisters on the lower lip 25 hours after it was suggested to him "while in hypnotic trance" that he appeared rundown and debilitated, he felt as if he were catching cold, and fever blisters were forming on his lower lip. Heilig and Hoff (1928) had previously demonstrated a similar effect in an experiment with three "neurotic" women. Their procedure was as follows: After a formal hypnotic induction, an intense emotional reaction was elicited from each *S* by suggesting an extremely unpleasant experience related to her previous life history. During the excitement, the experimenter stroked the *S*'s lower lip and suggested a feeling of itch such as she had experienced previously when a cold sore was forming. Within 48 hours after the termination of the experiment small blisters had appeared on the lower lip of each *S*. This report also includes the following data: at least two of the *S*s had a history of recurrent herpes labialis following emotional arousal; determination of the opsonic index before and after the hypnotic experiment indicated that the *S*s' physiological resistance was reduced after the experiment; herpetic blisters could not be induced when the hypnotized *S*s were given direct suggestions that such blisters were forming without at the same time eliciting an emotional reaction.

The above studies can be placed in broader context by noting the following: (*a*) The herpes simplex virus appears to be ubiquitous and ready to produce illness whenever the normal balance between it and the host is disturbed not only by fever, allergic reactions, sunburn, and so forth, but also by emotional stress and by symbolic stimulation which has significance for the person (Sulzberger & Zardens, 1948). (*b*) Some persons show recurrent attacks of herpes simplex in the same localized area (Veress, 1936); in some cases the attacks appear to be closely related to "emotional conflicts" or to stimulation which tends to elevate the level of "arousal" (Blank & Brody, 1950; Schneck, 1947). These findings suggest that an "hypnotic induction" procedure and specific suggestions of blister formation may not be necessary to induce herpetic blisters in appropriately predisposed persons. An experiment along the following lines is indicated: An experimental group consisting of persons with a history of herpes labialis should be given appropriate stimulation to induce emotional arousal *without* an hypnotic procedure. A second experimental group consisting of persons who do not have a history of herpes should be placed in "deep hypnosis" and given specific suggestions of cold sore formation. It can be hypothesized that some of the unhypnotized *S*s in the first group will show herpetic blisters within a day or so after the experiment. It would be of interest to determine if any of the "deeply hypnotized" *S*s in the second group will show cold sores after the experiment.

Induction of Localized (Nonherpetic) Blisters by Hypnotic Stimulation

Pattie (1941) has reviewed 11 experiments which ostensibly demonstrate that localized blisters (excluding cold sores) can be evoked by direct suggestions given to somnambulistic hypnotic *S*s. A relatively well controlled experiment reported by Hadfield (1917) can be taken as the prototype of these investigations: After the *S* was hypnotized, an assistant touched his arm while Hadfield gave continuous suggestions that a red-hot iron was being applied and that a blister would form in the burned area. The arm was then bound in a sealed bandage and the *S* was watched continuously during the following 24 hours. At the end of this period the bandage was opened in the presence of three physicians and, on the designated area, the beginning of a blister was noted which gradually developed during the day to form a large bleb surrounded by an area of inflammation. Although the other experiments followed this general pattern, there are numerous variations: in some instances, the experimenter stated that a blister would form after a definite time interval and in other instances no time was specified; some *S*s were instructed to awaken immediately after the suggestion of bulla formation and others were not given such instructions until it was determined if the blister had formed; although in most instances the blister formed in the area specified, in at least two instances (Jendrassik, 1888; Smirnoff, 1912) the bleb formed in another body area. Also, in at least two experiments (Rybalkin, 1890; von Krafft-Ebing, 1889, pp. 26–27, 58–59) the controls were not satisfactory; the *S*s were not observed during the intervening period and it is possible that they may have deliberately injured the area.

Two additional cases have been reported since the publication of Pattie's (1941) review. Ullman's (1947) *S*, mentioned in the preceding section of the present paper, had previously been cured of hysterical blindness and had previously shown herpetic blisters after hypnotic stimulation. In an additional hypnotic session, the *S* was induced to recall the battle in which he had recently participated and was given the suggestion that a small particle of molten shell fragment had glanced off the dorsum of his hand. At this point in the procedure, the experimenter brushed the hand with a small flat file to add emphasis to the suggestion. Pallor followed immediately in this circumscribed area approximately 1 centimeter in diameter; after 20 minutes a narrow red margin was evident about the area of pallor and after 1 hour the beginning of a blister was noticeable. The *S* was then dismissed and returned approximately 4 hours later; at this time a bleb about 1 centimeter in diameter was evident. (The *S* was not observed during the intervening period.) More recently, Borelli and and Geertz (Borelli, 1953) succeeded in inducing dermatological alterations which superficially resembled blister formation in a 27-year-old patient with "neurodermatitis." During "deep hypnosis" a coin was placed on the normal skin of the hand and it was suggested that a blister would form within a day at the spot were the fictitious burn was occurring. The next day the patient showed a sharply circumscribed and elevated area at the designated spot which superficially resembled a blister but could be more appropriately described as white dermographism.

With few if any exceptions investigators reporting positive results emphasize that they selected somnambulistic hypnotic *S*s for their experiments; however, a number of workers using similar procedures with somnambulistic *S*s have reported negative results in all cases (Sarbin, 1956; Wells, 1944), or have reported negative results with the majority of such *S*s and positive results only in rare cases (Hadfield, 1920). These negative findings appear to indicate that appropriate suggestions given to "deeply hypnotized" persons may be necessary but by no means sufficient conditions for this phenomenon.

An additional factor which appears necessary is indicated by the following. The 13 persons who gave ostensibly positive results were not a cross section of the normal population: prior to the experiment, one had been cured of hysterical blindness and one had been cured of hysterical aphonia; during the time of the experiment, six were diagnosed as hysterical and one was being treated for "shell-shock." At least five of these *S*s had histories of localized skin reactions: one had "a delicate skin" and showed labile vasomotor reactions (Doswald & Kreibich, 1906, Case 1), a second had suffered from "neurotic skin gangrene" and had a history of wheals following emotional arousal (Doswald & Kreibich, 1906, Case 2), a third had "a delicate skin" plus "dermographia of medium grade" (Heller & Schultz, 1909), a fourth had suffered from "hysterical ecchymoses" (Schindler, 1927), and a fifth was afflicted with atopic dermatitis (Borelli, 1953). This suggests that the induction of localized blisters by hypnotic stimulation may be possible only in a small group of persons with a unique physiological predisposition. What is the nature of this "predisposition"? The data summarized below suggest a tentative answer.

Blister formation and wheal formation apparently involve similar physiological and biochemical processes: the circular wheal of urticaria, the linear wheals of dermographism, and the blister resulting from a burn can be viewed as variations of the "triple response" of the skin to injury, consisting of the release of histamine or a histamine-like substance such as 5-hydroxytryptamine (serotonin) from the Mast cells, a local dilation of the minute vessels, an increase in permeability of the vessels, and a widespread arteriolar dilation (Lewis, 1927; Nilzén, 1947). Nearly every type of stimulus that produces whealing when applied to the skin will lead to blistering if rendered more intense, and blister formation appears to differ from wheal formation primarily in that the increased permeability of the vessel walls is of greater degree, the transuded fluid typically forms a pool in the superficial layers of the skin, and the epidermal layers are gradually forced asunder (Lewis, 1927). This close relationship between wheals and blisters appears to be significant because of the following:

1. In at least two of the "successful" hypnotic experiments (Borelli, 1953; Doswald & Kreibich, 1906, Case 2) the dermatological changes induced were much more similar to wheals that to blisters.

2. A critical reading of the other reports suggests that the histological findings were rarely so clear-cut as to definitely conclude that blisters and not wheals were produced.

3. Some unhypnotized persons show localized wheals when recalling former experiences in which such dermatological effects occurred.

4. Some unhypnotized persons show localized wheals after mild me-

chanical stimulation.

Moody (1946, 1948) has presented two case studies of patients who developed localized wheals when recalling former experiences in which wheals occurred. The first patient had been previously hospitalized for sleepwalking with aggressive behavior. On one occasion, during this earlier hospitalization, the patient's hands had been tied behind his back during sleep and wheals had formed in the traumatized area. At a later time, when recalling this experience after hexobarbital administration, wheals appeared on both forearms in the area which had previously been tied. On at least 30 occasions when recalling earlier experiences of physical injury, the second patient (who was being treated for "nervous breakdown") showed swelling, bruising, and bleeding in the body parts were the original injury presumably occurred; for instance, when remembering a former occasion when she had been struck across the dorsum of both hands with a cutting whip, the patient showed wheals on both hands in the respective areas. Along similar lines, Graff and Wallerstein (1954) reported that during a therapeutic interview a 27-year-old sailor, who had a tattoo of a dagger on his arm, suddenly showed a wheal reaction sharply limited to the outline of the dagger. The wheal subsided after this session but reappeared again in the same way during a subsequent interview. The authors interpret the patient's free associations as indicating that the wheal had symbolic significance for the patient. Brandt (1950) has reported similar cases of patients showing sharply localized wheal reactions which appeared to be closely related to symbolic stimulation.

Dermographism (that is, wheal formation in response to a single moderately strong stroking of the skin) is not as uncommon as is generally assumed. Testing 84 apparently normal young men, Lewis (1927) found a detectable swelling of the skin as a reaction to a single firm stroke in 25%; in 5% a full wheal developed. Some persons also show wheal formation at sites of mild pressure stimulation such as around a wristwatch strap, a belt, or a collar. Graham and Wolf (1950) reported an experimental study of 30 such persons who had a history of urticaria and showed "spontaneous" wheals in areas of mild pressure. All of these *Ss* also showed dermographism although in some this was not apparent until stressful interviews had altered the condition of the skin vessels. Skin temperature measurements and indirect measurements of the state of the minute vessels (reactive hyperemia threshold) indicated that the *Ss* were prone to respond with vasodilation of both arterioles and minute vessels to numerous stimuli. Since in all but one of the successful hypnotic experiments tactual stimulation was employed to localize the pseudotrauma and since in many of the experiments the stimulus object was a small piece of metal and was either allowed to remain in contact with the skin or was replaced by a bandage, it appears plausible, as Weitzenhoffer (1953, p. 144) has suggested, that similar physiological mechanisms may be responsible for the above types of urticaria factitia and for at least some cases of the hypnotic production of localized "blisters."

The above data suggest an experiment as follows: Persons who show gross vasomotor alterations during seemingly slight changes in the stimulating situation or who show dermographism under normal conditions or during stress should be given the following instructions *without* a preliminary "hypnotic induction"—"Try to visualize a blister (in a specified area) and tell yourself repeatedly that such a blister is forming." If the *S*s are adequately motivated to comply with these odd instructions, it can be hypothesized that some will show dermatological changes related to vesiculation. A second experimental group consisting of persons who do *not* show signs of vasomotor lability should be given suggestions of blister formation after an "hypnotic induction" procedure and when they appear to be in "the trance state." It would be of interest to determine if these "hypnotized" *S*s will show any cutaneous reactions which are involved in the formation of a blister.

Cure of Warts by "Hypnosis"

Since the genesis of warts appears to be causally related to virus activity and since present day methods of treating warts are "roundabout and nonspecific" (Pillsbury, Shelley, & Kligman, 1956, p. 690), recent reports indicating that appropriate suggestions given to a hypnotized person are singularly effective in curing these benign epitheliomas are of unique interest. Asher (1956) found that suggestions of wart disappearance given to 25 hypnotizable patients resulted, after 4 to 20 treatments, in a complete cure in 15, a marked decrease in the number of warts in 4, and no apparent change in 6 patients. In these cases the warts before treatment varied from 2 to 53 and were present from 3 months to 6 years. Eight unhypnotizable patients given similar suggestions showed no diminution in the number of warts; however, in these cases the treatment was discontinued after 10 sessions. In a more extensive investigation, Ullman and Dodek (1960) attempted to relieve warts by hypnotic suggestions in 62 adults attending an outpatient clinic. At weekly intervals each patient was given suggestions of sleep and drowsiness followed by suggestions to determine "the depth of hypnosis"; when the patient was judged to be at "the period of maximum hypnotic effect," he was told that the warts would begin to disappear. Of the 47 patients rated as "poor hypnotic subjects," only 2 showed wart regression within a 4-week period. However, 6 of the 15 patients rated as "good hypnotic subjects" had been cured of multiple common warts (or, in one case, of a single common wart) within 2 weeks following the initiation of treatment; within a 4-week period, 8 of the 15 showed wart involution. In these successful cases the mean duration of the warts prior to treatment was 19 months with a range of 3 weeks to 6 years.

The above investigations are open to the criticism that the warts may have shown spontaneous involution within the same period of time if no hypnotic treatment had been given. A recent study, however, appears to have satisfactorily controlled this factor. After an "hypnotic induction" consisting of eye fixation and suggestions of relaxation, Sinclair-Gieben and Chalmers (1959) suggested to 14 patients (with common warts present bilaterally for at least 6 months) that the warts on *one* side of the body would disappear. Ten of the 14 patients showed "adequate depth of hypnosis" as indicated by compliance with a simple posthypnotic suggestion and by partial or complete amnesia. Within 5 weeks to 3 months, 9 of these 10 "hypnotizable" patients showed wart involution on the "treated" side while the warts on the "control" side remained unchanged. (In one patient the "untreated" side showed wart regression 6 weeks after the "treated" side had been cured.) No benefit was observed from this treatment in the four patients who were not able to attain "adequate hypnotic depth."

Although the above studies indicate that symbolic stimulation is effective in inducing wart involution in some *S*s who are able to attain "a deep hypnotic state," equally successful results have been reported for a variety of suggestive procedures which do not involve an "hypnotic induction" or "the trance state." Grumach (1927) found that 16 of 18 patients with longstanding warts showed complete regression of these structures within 1 to 4 months after being given, at intervals of 8 to 14 days, an intramuscular placebo injection (normal saline) in the upper arm while, at the same time, being told that they were receiving a new and powerful wart remedy. Allington (1934) followed-up 84 patients with longstanding warts treated with an intragluteal placebo injection (distilled water); 35 (or 41.7%) were relieved of plane warts or common warts after only one injection, 4 were cured after two injections, and 1 after three injections. Bloch (1927) reported comparable results with a somewhat different procedure. The patient was blindfolded and his hand was placed on a table containing an electric apparatus; although the electricity was started no current reached the patient. The warts were then painted with an innocuous dye, the blindfold was removed, and the patient, now confronted with the luridly colored warts, was told that the warts were dead and must not be washed until they had disappeared. Of 179 patients thus treated and adequately followed-up, 55 (or 30.7%) showed wart involution after the first session and an additional 43 patients (or 24%) showed wart involution after additional session extending over a period varying from 1 week to 3 months. Using similar procedures, Bonjour (1929), Sulzberger and Wolf (1934), and Vollmer (1946) reported success in a comparable percentage of cases with warts of from 2 to 6 years duration. In general, these suggestive procedures were more effective when the patient showed multiple warts rather than a single wart and when the warts were of the juvenile type rather than the common type; this type of treatment also tended to be more successful with recent lesions and with younger patients.

Would a similar percentage of patients have shown spontaneous remission of warts if they had not been "treated" in the specified period of time involved in the above experiments? Memmescheimer and Eisenlohr (1931) matched 70 patients treated by a suggestive procedure—painting the warts with methylene blue and suggesting their disappearance—with 70 patients with similar warts of similar duration not given any treatment. The results were as follows: at the end of 1 month, 11 of the treated patients showed wart resolution as compared to only 2 of the patients in the control group; at the end of 3 months, 14 of the treated patients were cured as compared to only 5 of the untreated; however, at the end of 6 months, 20 patients in the control group showed wart involution as compared to only 17 patients in the treated group. The conclusion suggested by this study, namely, that suggestive treatment may accelerate a spontaneous physiological process leading to wart involution, is supported by additional investigations summarized below.

Similar physiological processes have been demonstrated when warts heal spontaneously and when they are cured in apparent response to symbolic stimulation. Unna (quoted by Samek, 1931) observed histologically that during spontaneous remission the normal cutis surrounding the wart showed a distinctive reaction consisting of hyperemia and cell proliferation. Other workers (Allington, 1952; Biberstein, 1944; Sulzberger & Wolf, 1934; Vollmer, 1946) have also noted a distinct inflammatory reaction immediately before spontaneous healing or before wart disappearance in apparent response to suggestion or to chemical treatment. In histological studies of warts undergoing involution in a patient treated by a suggestive procedure, Samek (1931) demonstrated a specific inflammatory reaction in the dermis consisting of dilation of blood vessels, hyperemia, edema, and perivascular infiltration of leucocytes (especially lymphocytes). Concomitant with this inflammatory reaction, mitoses became less frequent in the germinative epidermis (stratum mucosum); with mitoses almost at a standstill, the prickle-cell layer became thin, a normal stratum granulosum reformed, and the degenerated cells flaked off.

After a careful review of the above and related studies, Allington (1952) concluded that "at times the balance between susceptibility and immunity in warts must be a delicate one [and] only a slight shift may be needed to cause their disappearance." Vollmer (1946) had similarly concluded from an earlier review that a labile equilibrium must exist between the physiological processes which maintain the wart and those which cause wart involution and that appropriate verbal stimulation may alter the equilibrium in the direction of wart resolution by causing hyperemia in the surrounding tissue. A number of earlier workers (Sulzberger & Wolf, 1934; Zwick, 1932) had also pointed to vasomotor changes as crucial factors in wart remission and, more recently, Ullman (1959) presented data indicating that when warts are treated by suggestion an "affective response" is induced in the patient and the mechanism of healing may be dependent on local vascular alterations which accompany the emotional reaction. Since a number of investigations reviewed in an earlier section of the present paper suggest that localized vasodilation and localized vasoconstriction can be induced in *some* individuals by symbolic stimulation—e.g., by asking the individual to recall former experiences in which such vasomotor alterations occurred (Menzies, 1941)—further investigations are required to determine the following: (*a*) Are local vasomotor changes consistently present when wart resolution is occurring after suggestive treatment? (*b*) If so, do such vasomotor effects accelerate a natural physiological process of wart remission? (*c*) Is treatment of warts by suggestive procedures relatively more successful in persons who show vasomotor lability, that is, in persons who respond with a greater than average degree of vasodilation or vasoconstriction to symbolic stimulation or to emotion-inducing stimulation?

The Physiological Correlates of "The Hypnotic State"

The studies reviewed above suggest the general conclusion that many if not all of the physiological effects which can be induced in some *S*s during "hypnosis" can also be induced in some persons without hypnosis. The experiments reviewed below suggest that it is difficult if not impossible to find a physiological index which differentiates "the hypnotic state" from "the normal waking state."

During recent years an extensive number of experiments have been designed to determine if "hypnosis" is characterized by an elevated or depressed metabolic rate, heart rate, blood pressure, skin conductance, respiratory rate, digital blood volume, etc. All of these investigations lead to a similar conclusion: Physiological functions vary in the same way during "hypnosis" as they do during "waking" behavior. Taking energy expenditure as the example, the evidence indicates that metabolic rate may be elevated, may be depressed, or may not be significantly altered during "the hypnotic state": Grafe and Mayer (1923) found that hypnotized *S*s tended to show an elevated metabolic rate; von Eiff (1950) found that 16 *S*s showed an average depression of 7% in "basal" metabolic rate during hypnosis; and Whitehorn, Lundholm, Fox, and Benedict (1932) reported that oxygen consumption was not significantly affected by hypnosis. Since the metabolic rate is elevated during "emotional arousal" and is depressed during relaxation and sleep (Best & Taylor, 1950, p. 622), these results are only superficially contradictory: Experimenters finding that "hypnosis" depresses metabolism (von Eiff, 1950) had instructed their *S*s to become relaxed, drowsy, and sleepy and had not given additional suggestions that could lead to arousal; investigators reporting that "the hypnotic state" does not affect metabolism (Whitehorn et al., 1932) had trained their *S*s over a period of days to insure maximal relaxation when the metabolic rate was determined during the control experiment; experimenters finding that heat production was elevated during "hypnosis" (Grafe & Mayer, 1923) had activated the *S*s by suggesting various emotional experiences.

Investigations designed to determine if "hypnosis" is characterized by an elevated or depressed level of skin conductance have produced comparable results; during "the hypnotic trance" *S*s may show an elevation, a slight decrease, or no significant change in palmar conductance (Barber & Coules, 1959; Davis & Kantor, 1935; Estabrooks, 1930; Levine, 1930). Since an elevated conductance level generally indicates an elevated "activation" level and a low level of conductance generally indicates a low level of "arousal" (Duffy, 1957; Woodworth & Schlosberg, 1954), these results are in agreement as follows: (*a*) Hypnotic *S*s show an elevated level of palmar conductance when they *carry out* suggestions which involve effort or activity (Barber & Coules, 1959; Davis & Kantor, 1935). (*b*) When given suggestions of relaxation and drowsiness, *S*s participating in hypnotic experiments may show a decrease or no significant change in palmar conductance; if the *S* accepts the suggestions literally and relaxes, he shows a fall in conductance (Davis & Kantor, 1935; Estabrooks, 1930; Levine, 1930); if the *S* is aware that suggestions of drowsiness and relaxation are not meant to be taken literally, i.e., if he has learned from previous participation in hypnotic experiments that to carry out subsequent suggestions properly he must remain alert, he generally shows no significant change in conductance (Barber & Coules, 1959). Investigations along similar lines which support the general conclusion that the "hypnotized" person does not differ significantly from the normal person in heart rate, respiratory rate, blood pressure, digital blood volume, etc. have been reviewed by Gorton (1949) Weitzenhoffer (1953), Sarbin (1956), and Crasilneck and Hall (1959).

Some years ago it seemed that the electroencephalograph would prove to be a valuable tool for determining when a person was or was not "hypnotized." This hope has not been realized. Extensive work in this area, reviewed by Weitzenhoffer (1953) and Chertok and Kramarz (1959), has demonstrated that in the great majority of instances the hypnotized person continues to show his characteristic waking pattern on the EEG. However, if the operator makes it clear to the *S* that he should actually sleep—for example, by not disturbing the *S* after instructing him to sleep—some *S*s participating in hypnotic experiments show delta activity on the EEG, indicating that they have literally gone to sleep (Barker & Burgwin, 1948; Schwarz et al., 1955), and others show "periods of brief flattening out of the record . . . sometimes accompanied by infrequent isolated theta rhythms," indicating that they have gone into a light sleep (Chertok & Kramarz, 1959, p. 233). However, when the *S* is once more stimulated verbally by the hypnotist, he again shows his characteristic waking pattern on the EEG. In brief, studies employing the EEG indicate that the "hypnotized" person remains normally awake until it is made clear that he should literally go to sleep and is then permitted to sleep.

Within recent years, Lovett Doust (1953) and Ravitz (1951, 1959) have proposed two additional physiological indices of "the hypnotic state." Employing three hysterics and one psychopath as *S*s, Lovett Doust found that "the induction of hypnosis" was consistently accompanied by a significant fall in arterial oxygen saturation levels as measured by discontinuous spectroscopic oximetry at the fingernail fold. However, the term "induction of hypnosis," as used in this report, does *not* imply that the *S*s carried out one or more of the classical hypnotic behaviors, e.g., limb rigidities, negative or positive hallucinations; on the contrary, by this term the author refers to no more than the following: After being given suggestions of drowsiness, lethargy, and sleep, the *S*s appeared passive and lethargic. Since a person who appears passive and lethargic is not necessarily "hypnotized" (that is, does not necessarily carry out any of the classical hypnotic behaviors) and since a person who carries out all of the classical hypnotic behaviors does not necessarily appear drowsy or passive (Barber, 1960b; Barber & Coules, 1959; Wells, 1924), Lovett Doust's findings are open to the following interpretation: A relative anoxemia is found during drowsiness or passivity and is not necessarily found when a person is "in the hypnotic state," i.e., when he carries out the classical hypnotic behaviors. Supporting evidence for this interpretation is presented in a previous study by the same investigator (Lovett Doust & Schneider, 1952) which demonstrated a similar fall in oximetric values during sleep.

Measuring standing potentials between the forehead and the palm of the hand, Ravitz (1951, 1959) found that the "hypnotic induction" procedure was accompanied by either a gradual increase or decrease in mean potential and "the trance state itself, following induction" was typically characterized by a voltage decrease and by an increased regularity of the direct current (DC) tracings. However, additional data presented in the reports suggest that a decrease in voltage and an increased regularity of the DC tracings may be present whenever a person is relaxed and shows a low "arousal" level; for example, Ravitz notes that a decrease in voltage and an increase in regularity of the tracings are found during sleep and that increased voltage and decreased regularity are found during "changes in energy level," excitability, loquaciousness, grief, anxiety, and so forth. Since, as pointed out above and as will be discussed further below, a *S* need not show relaxation or passivity when he carries out the classical hypnotic behaviors, no conclusions can be deduced from these findings until the following hypotheses are experimentally confirmed or disproved: (*a*) Unhypnotized *S*s, i.e., *S*s who do *not* carry out such behaviors as limb rigidity, negative and positive hallucinations, age-regression, or amnesia when given appropriate suggestions, show a relative decrease in voltage and an increased regularity of the DC tracings when instructed to become relaxed and passive. (*b*) If an "hypnotic induction" leading to drowsiness and passivity is not employed, if, on the contrary, a direct suggestive procedure is used as described by Wells (1924) and Barber (1960b), "deeply hypnotized" *S*s, i.e., *S*s who carry out all of the classical hypnotic behaviors, do *not* show the above indices of "the trance state."

The above investigations and more recent speculations concerning the neurophysiological correlates of "hypnosis" (Arnold, 1959; Roberts, 1960; West, 1960) appear to be based on the following implicit assumptions: (*a*) When a person carries out the type of behavior which has been historically associated with the term "hypnosis" he is in "an altered state" from his normal self, specifically, in "a trance state" or "an hypnotic state." (*b*) This "altered state" is of such a kind as to include a distinct and consistent type of physiological functioning which is rarely if ever present when a person is not carrying out hypnotic behavior. Although these assumptions are by no means limited to recent investigations (they are present in many if not all theories of "hypnosis" since the days of Mesmer), the evidence summarized below suggests that they are open to question.

Hypnotic Behavior without an "Hypnotic Induction"

Since *S*s participating in hypnotic experiments are almost always given an "hypnotic induction" consisting of suggestions of relaxation, drowsiness, and sleep, and since such a procedure is generally effective in inducing an appearance of lethargy or "trance," it often seems as if hypnotic behavior is a function of, or closely related to, "the trance state." However, in a pioneering study, Wells (1924) demonstrated that direct commands (e.g., "Your arm is insensitive to pain," "You cannot speak your name"), repeated emphatically for a few seconds, were sufficient to elicit anesthesia, amnesia for name, limb rigidity, hallucinatory pain, total amnesia, automatic writing, and posthypnotic behavior in a large proportion of male college students. Wells insisted that his *S*s did not appear relaxed, drowsy, or lethargic and that he obtained results more quickly and with a larger proportion of *S*s by such a direct procedure than by an "hypnotic induction" designed to induce "trance."

Recent investigations appear to confirm Wells' results. In one study (Barber, 1960b) a female student research assistant (untrained as a hypnotist) gave 236 students at a girl's college direct suggestions (each suggestion requiring either 30 or 45 seconds) of body immobility—"Your body is heavy, rigid, solid; it's impossible for you to stand up; try, you can't,"—arm heaviness, arm levitation, hand rigidity, inability to say name, hallucination of thirst, selective amnesia, and posthypnotic behavior. Although an "hypnotic induction" was not employed, 49 *S*s (or 20.8%) immediately carried out at least six of the eight suggestions and a total of 109 *S*s (or 46%) carried out at least half of the suggestions. The postexperimental reports of these *S*s were indistinguishable from the reports of persons who are said to be "hypnotized," e.g., "I just couldn't get up from the chair," "I was amazed when I couldn't speak my name," "I felt I was dying from thirst." In another study (Barber, 1960b, 1960c) the results of such a direct procedure were compared with the results of a formal "hypnotic induction" procedure. In the first part of this experiment 70 attendants, nurses, and clerical workers at a state hospital (who agreed to participate in an experiment on "imagination") were given a series of suggestions (each suggestion requiring 30 seconds) appropriate to induce arm rigidity, arm levitation, limb heaviness, limb anesthesia, hallucinations of thirst, heat, and cold, eye catalepsy, and hypnotic dream. Similar results were obtained as in the above study: 20 *S*s (or 28.6%) immediately carried out at least seven of the nine suggestions and a total of 34 *S*s (or 48.6%) carried out at least five of the suggestions. In the second part of this experiment the same *S*s were given an "hypnotic induction" procedure (consisting of suggestions of relaxation, drowsiness, and sleep) and then given the suggestions of arm rigidity, arm levitation, limb anesthesia, etc., as in the preceding experiment. Although the *S*s now appeared to be "in trance" (and stated, after the experiment, that they had "felt hypnotized") a high correlation ($r = .84$) was obtained between scores in the two sessions; in general, *S*s who carried out one or two suggestions in the first part of the experiment carried out the same one or two suggestions after the "hypnotic induction" and *S*s who responded positively to all of the suggestions in the second part of the experiment had also carried out all of the suggestions without the "hypnotic induction."

Related to the above are the results of other recent investigations (Barber 1958a; Fisher, 1954) which indicate the following:

1. If *S*s participating in "hypnosis" experiments show lethargy, drowsiness, or other signs of "trance," these characteristics can be readily removed and the "good" *S*s will continue to carry out the hypnotic performances if instructed: "Be perfectly awake. Come out of 'trance' but continue to obey my commands."

2. Many if not all "good" hypnotic *S*s carry out all suggestions given during the posthypnotic period, i.e., after they are told to wake up, as long as they believe that their relationship with the operator remains that of subject and hypnotist.

In brief: Investigations which propose to find the physiological correlates of "hypnosis" uncritically assume that hypnotic behavior is a function of "the trance state"; this assumption is open to question. Appropriately predisposed persons do not need an "hypnotic induction" and need not appear to be in "trance" to carry out many if not all of the behaviors which have been associated with the term "hypnosis."

Summary and Conclusions

1. The normal person who is asked to "concentrate away from" red and green responds to the Ishihara in the manner characteristic of the hypnotic "color-blind" subject. An "hypnotic induction" procedure and "the trance state" may also be superfluous to eliciting the behavior which characterizes "hypnotic blind" or "hypnotic deaf" subjects; the evidence reviewed suggests that similar performances can be induced in normal persons by simply instructing them to remain inattentive and unresponsive to visual or auditory stimuli.

2. A number of physiological effects which have been considered as peculiar to "the hypnotic state" appear to be relatively commonplace performances; e.g., although suggestions of eating a delicious meal are at times effective in evoking gastric and pancreatic secretions and in inhibiting gastric hunger contractions in some "deeply hypnotized" subjects, it is not uncommon for normal persons to show similar gastrointestinal effects when they visualize the ingestion of savory food.

3. A group of so-called "hypnotic" phenomena—production of localized blisters, cure of warts, alteration of blood glucose levels, production of tachycardia or cardiac block—can apparently be elicited with or without an "hypnotic induction" in a small number of individuals who possess a specific lability of the physiological systems involved.

4. An extensive series of experiments has failed to find a physiological index which differentiates "the hypnotic state" from "the waking state."

5. A series of experiments comparing the results of an "hypnotic induction" procedure with the results of a direct suggestive procedure indicate that appropriately predisposed persons do not need an "hypnotic induction" and need not appear to be in "the trance state" to carry out the typical behaviors which have been associated with the word "hypnosis."

6. Further investigations into the nature of "hypnosis" might well bypass the concepts of "hypnotic induction" and "trance state" and focus on biographical and situational factors which may account for certain individuals responding to symbolic stimulation from another person with so-called "hypnotic" behavior, whether primarily motor responses (e.g., limb rigidity, eye catalepsy) or primarily physiological responses (e.g., tachycardia, wart involution).

REFERENCES

Allington, H. V. Sulpharsphenamine in the treatment of warts. *Arch. Dermatol. Syphilol., NY*, 1934, **29**, 687–690.

Allington, H. V. Review of psychotherapy of warts. *Arch. Dermatol. Syphilol., NY*, 1952, **66**, 316–326.

Arnold, Magda B. Brain function in hynosis. *Int. J. clin. exp. Hypnosis*, 1959, **7**, 109–119.

Asher, R. Respectable hypnosis. *Brit. med. J.*, 1956, **1**, 309–313.

Barber, T. X. Hypnosis as perceptual-cognitive restructuring: II. "Post"-hypnotic behavior. *J. clin. exp. Hypnosis*, 1958, **6**, 10–20. (a)

Barber, T. X. Hypnosis as perceptual-cognitive restructuring: IV. "Negative hallucinations." *J. Psychol.*, 1958, **46**, 187–201. (b)

Barber, T. X. Toward a theory of pain: Relief of chronic pain by prefrontal leucotomy, opiates, placebos, and hypnosis. *Psychol. Bull.*, 1959, **56**, 430–460.

Barber, T. X. "Hypnosis," analgesia, and the placebo effect. *J. Amer. Med. Ass.*, 1960, **172**, 680–683. (a)

Barber, T. X. The necessary and sufficient conditions for hypnotic behavior. *Amer. J. clin. Hypnosis*, 1960, **3**, 31–42. (b)

Barber, T. X. The reality of hypnotic behavior and the fiction of "hypnosis." Paper read at American Psychological Association, Chicago, September 1960. (c)

Barber, T. X., & Coules, J. Electrical skin conductance and galvanic skin response during "hypnosis." *Int. J. clin. exp. Hypnosis*, 1959, **7**, 79–92.

Barber, T. X., & Deeley, D. C. Experimental evidence for a theory of hypnotic behavior: I. "Hypnotic color-blindness" without "hypnosis." *Int. J. clin. exp. Hypnosis*, 1961, **9**, 79–86.

Barker, W., & Burgwin, Susan. Brain wave patterns accompanying changes in sleep and wakefulness during hypnosis. *Psychosom. Med.*, 1948, **10**, 317–326.

Bauch, M. Beeinflussung des Diabetes mellitus durch psychophysche Entspannungsübungen. *Dtsch. Arch. klin. Med.*, 1935, **2**, 149–166.

Bennett, L. L., & Scott, N. E. The production of electrocardiographic abnormalities by suggestion under hypnosis: A case report. *Amer. Practit.*, 1949, **4**, 189–190.

Berman, R., Simonson, E., & Heron, W. Electrocardiographic effects associated with hypnotic suggestion in normal and coronary sclerotic individuals. *J. appl. Physiol.*, 1954, **7**, 89–92.

Best, C. H., & Taylor, N. B. *The physiological basis of medical practice.* (5th ed.) Baltimore: Williams & Wilkins. 1950.

Biberstein, H. Immunization therapy of warts. *Arch. Dermatol. Syphilol., NY*, 1944, **50**, 12–22.

Blank, H., & Brody, M. Recurrent herpes simplex: A psychiatric and laboratory study. *Psychosom. Med.*, 1950, **12**, 254–260.

Bloch, B. Ueber die Heilung der Warzen durch Suggestion. *Klin. Wschr.*, 1927, **6**, 2271–2275, 2320–2325.

Bonjour, J. Influence of the mind on the skin. *Brit. J. Dermatol.*, 1929, **41**, 324–326.

Borelli, S. Psychische Einflüsse und reactive Hauterscheinungen. *Münch. med. Wschr.*, 1953, **95**, 1078–1082.

Brandt, R. A tentative classification of psychological factors in the etiology of skin diseases. *J. invest. Dermatol.*, 1950, **14**, 81–90.

Bykov, K. M. *The cerebral cortex and the internal organs.* New York: Chemical Publishing, 1957.

Callahan, A., & Redlich, F. C. Electroencephalography and opthalmology. *Amer. J. Ophthal.*, 1946, **29**, 1522–1533.

Carlson, A. J. *The control of hunger in health and disease.* Chicago: Univer. Chicago Press, 1916.

Chapman, L. F., Goodell, Helen, & Wolff, H. G. Increased inflammatory reaction induced by central nervous system activity. *Trans. Ass. Amer. Physicians*, 1959, **72**, 84–109.

Chertok, L., & Kramarz, P. Hyponosis, sleep and electro-encephalography. *J. nerv. ment. Dis.*, 1959, **128**, 227–238.

Crasilneck, H. B., & Hall, J. A. Physiological changes associated with hypnosis: A review of the literature since 1948. *Int. J. clin. exp. Hypnosis*, 1959, **7**, 9–50.

Crede, R. H., Chivers, N. C., & Shapiro, A. P. Electrocardiographic abnormalities associated with emotional disturbances. *Psychosom. Med.*, 1951, **13**, 277–288.

Davis, R. C., & Kantor, J. R. Skin resistance during hypnotic states. *J. gen. Psychol.*, 1935, **13**, 62–81.

Delhougne, R., & Hansen, K. Die suggestive Beeinflussbarkeit der Magen- und Pankreassekretion in der Hypnose. *Dtsch. Archiv. klin. Med.*, 1927, **157**, 20–35.

Dorcus, R. M. Modification by suggestion of some vestibular and visual responses. *Amer. J. Psychol.*, 1937, **49**, 82–87.

Doswald, D. C., & Kreibich, K. Zur Frage der posthypnotischen Hautphänomene. *Mh. prakt. Dermatol.*, 1906, **43**, 634–640.

Duffy, Elizabeth. The psychological significance of the concept of "arousal" or "activation." *Psychol. Rev.*, 1957, **64**, 265–275.

Dunbar, Flanders. *Emotions and bodily changes.* (4th ed.) New York: Columbia Univer. Press, 1954.

Dynes, J. B. An experimental study in hypnotic anesthesia. *J. abnorm. soc. Psychol.*, 1932, **27**, 79–88.

Erickson, M. H. A study of clinical and experimental findings on hypnotic deafness: I. Clinical experimentation and findings. *J. gen. Psychol.*, 1938, **19**, 127–150. (a)

Erickson, M. H. A study of clinical and experimental findings on hypnotic deafness: II. Experimental findings with a conditioned response technique. *J. gen. Psychol.*, 1938, **19**, 151–167. (b)

Erickson, M. H. The induction of color blindness by a technique of hypnotic suggestion. *J. gen. Psychol.*, 1939, **20**, 61–89.

Estabrooks, G. H. The psychogalvanic reflex in hypnosis. *J. gen. Psychol.*, 1930, **3**, 150–157.

Favill, J., & White, P. D. Voluntary acceleration of the rate of the heart beat. *Heart*, 1917, **6**, 175–188.

Fisher, S. The role of expectancy in the performance of posthypnotic behavior. *J. abnorm. soc. Psychol.*, 1954, **49**, 503–507.

Fisher, V. E. Hypnotic suggestion and the conditioned reflex. *J. exp. Psychol.*, 1932, **15**, 212–217.

Ford, W. L., & Yeager, C. L. Changes in electroencephalogram in subjects under hypnosis. *Dis. nerv. Syst.*, 1948, **9**, 190–192.

Gerard, R. W. General discussion of symposium. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior: The Hixon symposium.* New York: Wiley, 1951. P. 94.

Gigon, A., Aigner, E., & Brauch, W. Ueber den Einfluss der Psyche auf körperliche Vorgänge: Hypnose und Blutzucker. *Schweiz. med. Wschr.*, 1926, **56**, 749–750.

Gorton, B. E. The physiology of hypnosis. *Psychiat. Quart.*, 1949, **23**, 317–343, 457–485.

Grafe, E., & Mayer, L. Ueber den Einfluss der Affekte auf den Gesamtstoffwechsel. *Z. ges. Neurol. Psychiat.*, 1923, **86**, 247–253.

Graff, N. I., & Wallerstein, R. S. Unusual wheal reaction in a tattoo. *Psychosom. Med.*, 1954, **16**, 505–515.

Graham, D. T., & Wolf, S. Pathogenesis of urticaria: Experimental study of life situations, emotions, and cutaneous vascular reactions. *J. Amer. Med. Ass.*, 1950, **143**, 1396–1402.

Grether, W. F. A comment on "The induction of color blindness by a technique of hypnotic suggestion." *J. gen. Psychol.*, 1940, **23**, 207–210.

Grumach, L. Ueber Suggestivbehandlung von Warzen. *Münch. med. Wschr.*, 1927, **74**, 1093–1094.

Hadfield, J. A. The influence of hypnotic suggestion on inflammatory conditions. *Lancet*, 1917, **2**, 678–679.

Hadfield, J. A. The influence of suggestion on body temperature. *Lancet*, 1920, **2**, 68–69.

Harriman, P. L. Hypnotic induction of color vision anomalies: I. The use of the Ishihara and the Jensen tests to verify the acceptance of suggested color blindness. *J. gen. Psychol.*, 1942, **27**, 289–298.

Heilig, R., & Hoff, H. Ueber Psychogene Entstehung des Herpes labialis. *Med. Klin.*, 1928, **24**, 1472.

Heller, F., & Schultz, J. H. Ueber einen Fall hypnotisch erzeugter Blasenbildung. *Münch. med. Wschr.*, 1909, **56**, 2112.

Hernández-Peón, R., & Donoso, M. Influence of attention and suggestion upon subcortical evoked electrical activity in the human brain. In L. van Bogaert & J. Radermecker (Eds.), *First International Congress of Neurological Sciences.* Vol. III. London: Pergamon, 1959. Pp. 385–396.

Heyer, G. R. Psychogene Functionsstörungen des Verdauungstraktes. In O. Schwarz (Ed.), *Psychogenese und Psychotherapie körperlicher Symptome.* Wien: Springer, 1925. Pp. 229–257.

Hilgard, E. R., & Marquis, D. G. *Conditioning and learning.* New York: Appleton-Century, 1940.

Hinkle, L. E., Jr., & Wolf, S. A summary of experimental evidence relating life stress to diabetes mellitus. *J. Mt. Sinai Hosp.*, 1953, **19**, 537–570.

Huttenlocher, J., & Westcott, M. R. Some empirical relationships between respiratory activity and heart rate. *Amer. Psychologist*, 1957, **12**, 414. (Abstract)

Ikemi, Y., Akagi, M., Maeda, J., Fukumoto, T., Kawate, K., Hirakawa, K., Gondo, S., Nakagawa, T., Honda, T., Sakamoto, A., & Kumagai, M. Hypnotic experiments on the psychosomatic aspects of gastrointestinal disorders. *Int. J. clin. exp. Hypnosis*, 1959, **7**, 139–150.

Jasper, H. H., & Carmichael, L. Electrical potentials from the intact human brain. *Science*, 1935, **81**, 51–53.

Jasper, H. H., & Cruikshank, Ruth M. Electro-encephalography: II. Visual stimulation and the after-image as affecting the occipital alpha rhythm. *J. gen. Psychol.*, 1937, **17**, 29–48.

Jendrassik, E. Einiges über suggestion. *Neurol. Zbl.*, 1888, **7**, 281–283, 321–330.

Jenness, A., & Wible, C. L. Respiration and heart action in sleep and hypnosis. *J. gen. Psychol.*, 1937, **16**, 197–222.

King, J. T., Jr. An instance of voluntary acceleration of the pulse. *Bull. Johns Hopkins. Hosp.*, 1920, **31**, 303–304.

Kline, M. V., Guze, H., & Haggerty, A. D. An experimental study of the nature of hypnotic deafness: Effects of delayed speech feedback. *J. clin. exp. Hypnosis*, 1954, **2**, 145–156.

Koehler, M. Ueber die willkürliche Beschleunigung des Herzschlages beim Menschen. *Arch. ges. Physiol.*, 1914, **158**, 579–622.

Landis, C., & Slight, D. Studies of emotional reactions: VI. Cardiac responses. *J. gen. Psychol.*, 1929, **2**, 413–420.

Levine, M. Electrical skin resistance during hypnosis. *Arch. Neurol. Psychiat.*, 1930, **24**, 937–942.

Lewis, J. H., & Sarbin, T. R. Studies in psychosomatics: I. The influence of hypnotic stimulation on gastric hunger contractions. *Psychosom. Med.*, 1943, **5**, 125–131.

Lewis, T. *The blood vessels of the human skin and their responses.* London: Shaw, 1927.

Ljung, O. Ekg-Koronarinsuffizienz bei vegetativer Labilität. *Cardiologia*, 1949, **14**, 191–218.

LOFTUS, T. A., GOLD, H., & DIETHELM, O. Cardiac changes in the presence of intense emotion. *Amer. J. Psychiat.*, 1945, **101**, 697–698.

LOOMIS, A. L., HARVEY, E. N., & HOBART, G. Electrical potentials of the human brain. *J. exp. Psychol.*, 1936, **19**, 249–279.

LOVETT DOUST, J. W. Studies on the physiology of awareness: Oximetric analysis of emotion and the differential planes of consciousness seen in hypnosis. *J. clin. exp. Psychopathol.*, 1953, **14**, 113–126.

LOVETT DOUST, J. W., & SCHNEIDER, R. A. Studies on the physiology of awareness: Anoxia and the levels of sleep. *Brit. med. J.*, 1952, **1**, 449–453.

LUCKHARDT, A. B., & JOHNSTON, R. L. Studies in gastric secretions: I. The psychic secretion of gastric juice under hypnosis. *Amer. J. Physiol.*, 1924, **70**, 174–182.

LUNDHOLM, H. An experimental study of functional anesthesias as induced by suggestions in hypnosis. *J. abnorm. soc. Psychol.*, 1928, **23**, 337–355.

LUNDHOLM, H., & LOWENBACH, H. Hypnosis and the alpha activity of the electroencephalogram. *Charac. Pers.*, 1942, **11**, 144–149.

MCCLURE, C. M. Cardiac arrest through volition. *Calif. Med.*, 1959, **90**, 440–441.

MCDOWELL, M. Hypnosis in dermatology. In J. M. Schneck (Ed.), *Hypnosis in modern medicine.* (2nd ed.) Springfield, Ill.: Charles C Thomas, 1959. Pp. 101–115.

MADOW, L. Cortical blindness. *J. Neuropathol.*, 1958, **17**, 324–332.

MAINZER, F., & KRAUSE, M. The influence of fear on the electrocardiogram. *Brit. heart J.*, 1940, **2**, 221–230.

MALMO, R. B., BOAG, T. J., & RAGINSKY, B. B. Electromyographic study of hypnotic deafness. *J. clin. exp. Hypnosis*, 1954, **2**, 305–317.

MALMO, R. B., DAVIS, J. F., & BARZA, S. Total hysterical deafness: An experimental case study. *J. Pers.*, 1952, **21**, 188–204.

MARCUS, H., & SAHLGREN, E. Untersuchungen über die Einwirkung der Hypnotischen Suggestion auf die Funktion des vegetativen Systemes. *Münch. med. Wschr.*, 1925, **72**, 381–382.

MEMMESHEIMER, A. M., & EISENLOHR, E. Untersuchungen über die Suggestivebehandlung der Warzen. *Dermatol. Z.*, 1931, **62**, 63–68.

MENZIES, R. Further studies of conditioned vasomotor responses in human subjects. *J. exp. Psychol.*, 1941, **29**, 457–482.

MILLER, R. J., BERGEIM, O., REHFUSS, M. E., & HAWK, P. B. Gastric response to food: X. The psychic secretion of gastric juice in normal men. *Amer. J. Physiol.*, 1920, **52**, 1–27.

MIRSKY, I. A. Emotional factors in the patient with diabetes mellitus. *Bull. Menninger Clin.*, 1948, **12**, 187–194.

MOHR, F. *Psychophysische Behandlungsmethoden.* Leipzig: Hirzel, 1925.

MOODY, R. L. Bodily changes during abreaction. *Lancet*, 1946, **2**, 934–935.

MOODY, R. L. Bodily changes during abreaction. *Lancet*, 1948, **1**, 964.

NIELSEN, O. J., & GEERT-JORGENSEN, E. Untersuchungen über die Einwirkung der hypnotischen Suggestion auf den Blutzucker bei Nichtdiabetikern. *Klin. Wschr.*, 1928, **7**, 1467–1468.

NILZÉN, A. Studies in histamine. *Acta dermat.-venereol., Stockh.*, 1947, **27**, Suppl. 17, 1–67.

OGDEN, E., & SHOCK, N. W. Voluntary hypercirculation. *Amer. J. med. Sci.*, 1939, **198**, 329–342.

PATTIE, F. A. A report of attempts to produce uniocular blindness by hypnotic suggestion. *Brit. J. med. Psychol.*, 1935, **15**, 230–241.

PATTIE, F. A. The production of blisters by hypnotic suggestions: A review. *J. abnorm. soc. Psychol.*, 1941, **36**, 62–72.

PATTIE, F. A. The genuineness of unilateral deafness produced by hypnosis. *Amer. J. Psychol.*, 1950, **63**, 84–86.

PEASE, E. A. Voluntary control of the heart. *Boston med. surg. J.*, 1889, **120**, 525–529.

PILLSBURY, D. M., SHELLEY, W. B., & KLIGMAN, A. M. *Dermatology.* Philadelphia: Saunders, 1956.

POVORINSKIJ, J. A., & FINNE, W. N. Der Wechsel des Zuckergehalts des Blutes unter dem Einfluss einer hypnotisch suggerierten Vorstellung. *Z. ges. Neurol. Psychiat.*, 1930, **129**, 135–146.

RAGINSKY, B. B. Temporary cardiac arrest induced under hypnosis. *Int. J. clin. exp. Hypnosis*, 1959, **7**, 53–68.

RAVITZ, L. J. Standing potential correlates of hypnosis and narcosis. *Arch. Neurol. Psychiat.*, 1951, **65**, 413–436.

RAVITZ, L. J. Application of the electrodynamic field theory in biology, psychiatry, medicine, and hypnosis. *Amer. J. clin. Hypnosis*, 1959, **1**, 135–150.

ROBERTS, D. R. An electrophysiological theory of hypnosis. *Int. J. clin. exp. Hypnosis*, 1960, **8**, 43–55.

RYBALKIN, J. Brûlure du second degré provoquée par suggestion. *Rev. Hypnot., Paris*, 1890, **4**, 361–362.

SAMEK, J. Zum wesen der Suggestiven Warzenheilung. *Dermatol. Wschr.*, 1931, **93**, 1853–1857.

Sarbin, T. R. Physiological effects of hypnotic stimulation. In R. M. Dorcus (Ed.), *Hypnosis and its therapeutic applications.* New York: McGraw-Hill, 1956. Ch. 4.

Scantlebury, R. E., & Patterson, T. L. Hunger motility in a hypnotized subject. *Quart. J. exp. Physiol.*, 1940, **30**, 347–358.

Schindler, R. *Nervensystem und spontane Blutunge.* Berlin: Karger, 1927.

Schneck, J. M. Psychogenic component in a case of herpes simplex. *Psychosom. Med.*, 1947, **9**, 62–64.

Schultz, J. H., & Luthe, W. *Autogenic training.* New York: Grune & Stratton, 1959.

Schwarz, B. E., Bickford, R. G., & Rasmussen, W. C. Hypnotic phenomena, including hypnotically activated seizures, studied with the electroencephalogram. *J. nerv. ment. Dis.*, 1955, **122**, 564–574.

Sinclair-Gieben, A. H. C., & Chalmers, D. Evaluation of treatment of warts by hypnosis. *Lancet*, 1959, **2**, 480–482.

Smirnoff, D. Zur Frage der durch hypnotische Suggestion hervorgerufenen vasomotorischen Störungen. *Z. Psychother. med. Psychol.*, 1912, **4**, 171–175.

Solovey, Galina, & Milechnin, A. Concerning the nature of hypnotic phenomena. *J. clin. exp. Hypnosis*, 1957, **5**, 67–76.

Sulzberger, M. B., & Wolf, J. The treatment of warts by suggestion. *Med. Rec., NY*, 1934, **140**, 552–557.

Sulzberger, M. B., & Zardens, S. H. Psychogenic factors in dermatologic disorders. *Med. Clin. N. Amer.*, 1948 (May), 669–685.

Tarchanoff, J. R. Ueber die willkürliche Acceleration der Herzschläge beim Menschen. *Arch. ges. Physiol.*, 1885, **35**, 109–137.

Taylor, N. B., & Cameron, H. G. Voluntary acceleration of the heart. *Amer. J. Physiol.*, 1922, **61**, 385–398.

Ullman, M. Herpes simplex and second degree burn induced under hypnosis. *Amer. J. Psychiat.*, 1947, **103**, 828–830.

Ullman, M. On the psyche and warts: I. Suggestions and warts: A review and comment. *Psychosom. Med.*, 1959, **21**, 473–488.

Ullman, M., & Dodek, Stephanie. On the psyche and warts: II. Hypnotic suggestion and warts. *Psychosom. Med.*, 1960, **22**, 68–76.

Van de Velde, T. H. Ueber willkürliche Vermehrung der Pulsfrequenz beim Menschen. *Arch. ges. Physiol.*, 1897, **66**, 232–240.

Van Pelt, S. J. The control of heart rate by hypnotic suggestion. In L. M. LeCron (Ed.), *Experimental hypnosis.* New York: Macmillan, 1954. Pp. 268–275.

Veress, F. V. Beiträge zur Pathogenese des Herpes simplex. *9th Int. Conv. Dermatologists*, 1936, **2**, 242.

Vollmer, H. Treatment of warts by suggestion. *Psychosom. Med.*, 1946, **8**, 138–142.

von Eiff, A. W. Ueber die Moglichkeit einer Grundumsatzenkung durch Psychische Beeinflussung. *Artzl. Forsch.*, 1950, **4**, 611.

von Krafft-Ebing, R. *Eine experimentelle Studie auf dem Gebiete des Hypnotismus.* (2nd ed.) Stuttgart: 1889.

Weiss, B. Electrocardiographic indices of emotional stress. *Amer. J. Psychiat.*, 1956, **113**, 348–351.

Weitzenhoffer, A. M. *Hypnotism: An objective study in suggestibility.* New York: Wiley, 1953.

Wells, W. R. Experiments in waking hypnosis for instructional purposes. *J. abnorm. soc. Psychol.*, 1924, **18**, 389–404.

Wells, W. R. The hypnotic treatment of the major symptoms of hysteria: A case study. *J. Psychol.*, 1944, **77**, 269–297.

West, H. F., & Savage, W. E. Voluntary acceleration of the heart beat. *Arch. intern. Med.*, 1918, **22**, 290–295.

West, L. J. Psychophysiology of hypnosis. *J. Amer. Med. Ass.*, 1960, **172**, 672–675.

Whitehorn, J. C., Lundholm, H., Fox, E. L., & Benedict, F. G. Metabolic rate in "hypnotic sleep." *New England J. Med.*, 1932, **206**, 777–781.

Wolf, S., & Wolff, H. G. *Human gastric function: An experimental study of a man and his stomach.* (2nd ed.) New York: Oxford Univer. Press, 1947.

Woodworth, R. S., & Schlosberg, H. *Experimental psychology.* (Rev. ed.) New York: Holt, 1954.

Yeager, C. L., & Larson, A. L. A study of alpha desynchronization in the electroencephalogram utilizing hypnosis. Paper presented at American Electroencephalographic Society, Santa Fe, October 1957.

Zwick, C. G. Hygiogenesis of warts disappearing without topical medication. *Arch. Dermatol. Syphilol., NY*, 1932, **25**, 508–521.

(Received May 19, 1960)

Psychological Bulletin
1961, Vol. 58, No. 5, 420–428

THE NUMBER CONCEPT: A PHYLOGENETIC REVIEW

FRANK WESLEY
Portland State College

The data concerning number concepts of animals which have been reported so far do not agree with the general relationship between position in the phylogenetic scale and behavior. Rensch and Altevogt (1953) working with an elephant and Hicks (1956) with monkeys had limited success in establishing a "threeness" concept, while Koehler (1943) and his collaborators (Arndt, 1939; Braun, 1952; Lögler, 1959; Marold, 1939; Sauter, 1952; Schiemann, 1939) obtained a "sevenness" level on several species of birds.

Salman (1943) reviewed the number capacities of animals and found inadequate controls which allowed operation of rhythmic cues in most studies reported prior to 1939. Salman and other reviewers (Honigman, 1942; Koehler, 1951; Thorpe, 1956) considered rhythmic cues and other extraneous variables very well controlled in the studies reported in 1939 and thereafter, but did not mention the omission of certain operating procedures considered standard in the United States literature. The bird studies, originally reported in German, have been uncritically accepted by Hicks (1956), Morgan (1956), Newman (1956), and other American writers. It will therefore be the purpose of this paper to reexamine in detail the methodology used in studies reported since 1939.

Definition

There is considerable agreement among most investigators in the definition of a number concept. An animal is usually required to solve a problem without the aid of immediate physical variables. External cues such as size, shape, color, brightness, tactile, odor, etc., as well as internal ones arising from rhythmic motor patterns or other visceral or kinesthetic feedback should either be absent or randomized from trial to trial, so that the numerosity of the stimulus constitutes the only constant variable. Experiments which were designed to include any of the above physical variables on a nonrandomized basis will be omitted from this review. Included, however, will be those in which a number concept is reported by the authors though immediate and constant cues could have been responsible for the observed behavior.

BIRDS

Arndt (1939) tested number ability with various tasks in nine pigeons which received an average of 4,000 trials. To prevent rhythmic sequences Arndt presented peas on a turntable, exposing one pea at a time. Delays from pea to pea were from 1 to 60 seconds. His pigeon "Blaugrau" mastered the pecking of five peas only and would not touch a sixth pea when it appeared in the open slot. Another pigeon "Grau" learned a fourness problem on its first trial. In a tube experiment Arndt dropped peas from behind a screen at intervals varying from 1 to 20 seconds. On the animal's side the pea fell into a cup-shaped receptacle at the end of the tube. With this method pigeon

"Braunweiss" responded correctly with 55% in a twoness problem during the last hundred of 915 trials. When subsequently trained for a threeness problem, it responded correctly on the first seven trials, which means that without any negative transfer from the previous problem it picked the now correct third pea. Such an initial and highly accurate response strongly suggests the presence of extraneous cues. Arndt, however, looked upon it as "progress in learning" not realizing that even the most optimal "learning to learn" situation requires some negative transfer. In another experimental arrangement Arndt employed lid-covered boxes on a turntable. Again, only one box appeared in an open slot at any one time. With a twoness task one wheat kernel or one pea was placed into each of two successively appearing boxes. One pigeon "Blauweiss" learned to open these two successive boxes, but would not open a third box. When the two baits were distributed within three boxes, (1, 0, 1) the second box being empty, "Blauweiss" exhibited immediate learning, opening now three boxes, and leaving the fourth one untouched. From this behavior Arndt concluded that the bird had not learned to open a certain number of boxes, but learned to eat a certain number of peas. Gradually, within 6,000 further trials it learned to take six peas out of six boxes, not opening the seventh box. During the above experiments Arndt noticed that the birds would usually remain at the slot of the apparatus after they had responded correctly, and would turn away only after the turntable turned to present the negative stimuli. Arndt tested the possibility of differential acceleration as an extraneous cue, which may have been possible, since the turntable was operated manually. He asked another experimenter to turn the table for 600 trials and observed no differential results when compared with data from his own manipulations. He failed to note that the other experimenter also knew the correct number, and that subjective acceleration cues may have remained constant from one experimenter to another. In his review Thorpe (1956) describes Arndt, among other experimenters, as having "adopted quite extraordinary precautions to avoid errors of the 'Clever Hans' type" (p. 344). But "Clever Hans" could also solve problems when given by another experimenter who knew the correct answer.

Arndt obtained 65% correct responses as the highest level of performance on sets of 100 trials during thousands of trials. Such a low level of mastery and the frequent absence of negative transfer do suggest extraneous cues with both the tube and turntable experiments. Auditory variables which could have arisen from the experimenter or from a rattling of baits in the boxes during the turning were not controlled. Olfaction received no attention and boxes were not baited beyond the desired number. Another extraneous variable in Arndt's methodology could have been the nonrandomization of the amount of food ingested. Since it can be assumed that most of the peas were of equal size, visceral feedback could have presented a constant and immediate stimulus, and the correct response could have been based on quantity of food rather than on a mediated numerical concept. Thus, a quantity of food, an odor, a noise, or a "subjective" turning speed may alone or in combination be responsible for the results observed. Arndt's methodology therefore, does not warrant the conclusion that behavior based on numerosity was exhibited.

Concurrently with Arndt, Marold (1939) tested several parakeets on simultaneous and successive tasks. One parakeet was trained to discriminate between groups of two and three kernels. No learning was exhibited within 500 trials. The bird, apparently, depended too strongly on figure aid and changed to a position habit whenever the figure aid was withdrawn. To break this position habit Marold allowed the bird to eat the negative group of kernels after a positive response and observed positive results on the sixth block of 100 trials. The correctness level, however, did not rise above 57% correct within 1,100 trials. On the successive task Marold used rows of kernels and required her birds to eat *x* kernels without touching the $x+1$ kernel. Marold's parakeet "Grün" was trained to eat two kernels from a row varying from three to seven kernels. The distance between the kernels was altered from .5 to 0 centimeter and with decreasing distance, decreasing accuracy was observed. At the end of 900 trials the bird responded 87% correct, but the percentage dropped to 44 on a subsequent block of 100 trials which involved a further decrease in spacing. In an additional block behavior resembling experimental neurosis was reported.

Throughout her experiments Marold reported large individual differences, but she concluded that these differences arose from individual differences in treatment and in "Einfühlung." Such a statement suggests that the birds did receive differential treatment intentionally or unintentionally, which may have accounted for some of the results observed. Marold's simultaneous discrimination task was not free from differential size cues. Likewise, extraneous distance cues were present during her successive task, resulting in experimental neurosis at zero distance. Such behavior resembles the inability to differentiate between cues immediately present in the environment. The presence of these and other extraneous cues makes it difficult to ascribe Marold's observation to numerical behavior alone.

Schieman (1939) used a new method for investigating the ability of birds to act successively to numbers. He confronted his jackdaws with a row of 10 covered dishes and required them to uncover their lids in sequence. Baits were differentially distributed according to a prearranged pattern, so that sometimes a dish would contain two or more baits and sometimes none. His birds had to uncover a different quantity of dishes to obtain *x* baits. Odor cues were not controlled since the dishes beyond the correct number which were not to be uncovered did not contain bait. One jackdaw "Blau" exhibited its upper limit of $x=6$ and performed with 65% correct during 886 trials. Another jackdaw learned within 1,000 trials to differentiate between the eating of two, three, four, and five baits.

Schiemann reported that performance was lower on a task which required the opening of *x* dishes rather than the eating of *x* baits. Schiemann did not vary the amount of food per trial, so that on any one number task this could have presented immediate cues from visceral feedback. Such a hypothesis could explain the high performance with "baits eaten" and the random performance with "dishes uncovered."

Schiemann attempted further a combination of successive and simultaneous number discrimination. He presented his jackdaw "Grün" a stimulus card containing either two or four dots. According to this sample-number "Grün" was subsequently required to peck the indicated number of baits from a plate. After 1,200 trials a 70% correct response was obtained in the last 100 block. Schiemann believed that this demonstrated a success in the ability to act out a previously seen number. It should, however, be noted that the size of the sample dots remained constant throughout this task. With stimuli differing in one physical dimension this task may be compared with the disjunctive RT experiments so need not be related to number concepts. If unknown samples were presented, Schiemann states, the jackdaws appeared to be "completely helpless."

Koehler (1943) worked intensively with a 9-year-old raven named "Jacob" which received a total of approximately 12,000 trials during 794 working hours. On a series of trials "Jacob" learned to discriminate successfully between piles of baits having the following ratios: 4:5, 4:6, 6:5, and 7:6 (the first number indicating the positive stimulus). After having mastered these tasks it was not possible to train "Jacob" to discriminate on a 5:6 problem, though several hundred trials were administered with and without punishment and interspersed with rest periods. A naive bird was likewise unsuccessful. Koehler noted in a later film of the experiment that his assistant had a tendency to place the positive group closer to the forward margin of the experimental board during the discrimination series. Koehler did not mention whether correction of this placement cue preceded the 5:6 problem, but if it did the failure could be explained. As in the famous case of Kinnebrook the assistant was replaced but manual placing of baits was continued.

The most difficult task which "Jacob" learned was a multiple choice task with a sample-indicator. A sample card was placed on the ground indicating the required number by means of irregular dots. Around it five covered dishes were placed with their lids showing irregular dots from numbers two to six. The dots on the positive lids differed in size, shape, and configuration from the ones on the sample. "Jacob" was able to obtain the reward of either grain, fruit, cheese, or meat when all aids were withdrawn and when the dots were replaced by irregular pieces of plasticine. The breaking and kneading of the plastic material as well as placing it on the lids was done manually, and again inadvertent cues arising from this manipulation should not be excluded in the evaluation of the obtained results. Odor control was likewise seriously lacking and initiated only after Tinbergen reported to Koehler that a jay was able to detect mealworms by odor. During the 481 trials which were presented with the irregular plasticine dots only five were partially odor controlled by baiting several dishes. "Jacob" responded correctly on all of these five trials.

Braun (1952) worked with three parrots to investigate some combination tasks. For positive reinforcement hempseed or cheese was used. Negative reinforcement consisted of punishment with a stick but was applied only when "absolutely necessary." At other instances during her experiments Braun made loud noises, threw a wet sponge, or pulled tail feathers as methods of negative reinforcement.

One of Braun's parrots performed on a dish-row problem during a sixness task with an average of 75% correct responses. It is difficult, however, to evaluate the results and ascribe them only to a numerical concept since the physical proximity which the experimenter must have maintained with her animals in order to administer the various methods of punishment could have presented a host of extraneous cues. Furthermore variables such as odor and amount of food ingested were not controlled.

Eight magpies were used as subjects by Sauter (1952) who repeated some of the above tasks. Her food dishes were baited and covered in the observation room and manually spaced 10–12 centimeters in the experimental room. Her rewards were a variety of foods, with type and amount remaining equal within one task. A scare apparatus which was rarely used served for negative reinforcement. She tested four magpies on a dish-row. *X* baits were distributed in 25 different ways into 10 dishes. Magpie "Prinz" accomplished an $x=3$ problem on this task, but showed no transfer effect when a row of ten mealworms replaced the row of 10 dishes. The upper limit, $x=7$, was reported to have been reached with magpie "Felix" performing at a level of 74% correct at the end of 100 trials.

Half of Sauter's bait distributions on the dish-row problem did not contain zero spacings and if odor cues, which were not controlled, were postulated, they could account for 50% of the correct responses. An additional 25% correct responses could be assumed if chance behavior occurred on 50% of the trials which did have zero spacings and therewith a possible odor control. Thus, the odor variable alone could explain the 74% correct level reported for "Felix" during the $x=7$ task. Odor remained likewise uncontrolled in Sauter's simultaneous experiments. Dishes were manually placed and the selection and production of the irregular pieces of plasticine did not follow a prescribed procedure to assure the nonoccurrence of the experimenter's unintentional cues.

Braun's parrot "Jako" was again used by Lögler (1959) to perform on 16,076 additional trials on various number combinations. One such task involved successive presentations of flashes of light which varied in number and served to indicate the correct number of baits to be chosen on a dish-row. Numbers up to seven could be acted out in this way. Stimulus generalization of numerosity was reported on lower numbers when the visual indicators were replaced by auditory ones. On a single dish-row problem "Jako" was successful in obtaining eight baits which were differentially distributed into 11 dishes. The bird reached significant results after performing on chance level for 600 trials. Though odor was not controlled its possible interaction on the successive problem is not likely since 4/5 of Lögler's bait distributions contained one or several zero spacings. But, again, the manual placing of all dishes, the nonrandomization of food quantities, and the occasional deviations from the intended methods of scaring could have contributed extraneous cues.

All of the above reported bird studies used variable and highly subjective types of negative reinforcement. Most experimenters designed a scare apparatus intended for uniform punishment, but abandoned it early in their experiments. Arndt (1939) changed over to bait withdrawal while Marold (1939) used blasts of air to blow away negative kernels but allowed them to be eaten in other instances. She also reported spraying of water and gypsum powder into the birds' faces. At times Schiemann (1939) "scared" the birds but darkened the room at other occasions. It is not clear in most of the above reported experiments how often and in which instances the various punishment methods were used. Honigman (1942) and Salman (1943) who reviewed the above experiments believed them well controlled and Thorpe (1956) more recently termed them "technically beyond reproach" (p. 349) and stated that the use of the punishment apparatus made it impossible for the experimenter to give inadvertent signs. He does not mention the abandonment of the apparatus and the substitute method reported by the various authors.

Another serious lack common to the above bird studies is the absence of odor controls. Only Koehler (1943) controlled it partially in 37 trials out of an approximate total of 55,000 trials given by the above experimenters who offered such variable baits as flour, bread, seeds, fruits, cooked and raw meats, cheese, and others. The general assumption that birds are insensitive to odor may be quite fallacious. In a well controlled experiment Zahn (1933) found odor sensitivity equal and surpassing human thresholds on five different odors in experiments with pigeons, blackbirds, blue titmouse, robins, and hedgesparrows.

Manual placing of turntables, cups, lids, baits, dots, or plasticine was also present in all of the above bird studies. It was reported by Arndt (1939) and Koehler (1943) to have influenced their results on certain occasions. Elimination of these and other cues could have been assured only by complete mechanical presentation of the stimulus components, a methodology not adopted by any of the experimenters.

Fish

The counting capacities of minnows, stickleback, and other small fish were investigated by Rossmann (1959) who found that innate preferences of bait size, stimulus density, and motility interfered with numerosity throughout prolonged training on the simultaneous discrimination task. One motor act and one tonal quantity, however, could eventually be differentiated from two acts or two tones on the successive task. One minnow, e.g., required a sequence of 170 negative reinforcement trials before it learned to eat the first bait without touching the second. The experiment was well controlled in regard to odor, bait-size, and rhythm. Training to numbers above one could not be established and Rossmann concluded that a number concept in fish can therefore not be postulated.

Mammals

Rodents

Hassmann (1952) experimented with 13 squirrels employing the methodology used by Koehler and his collaborators with the bird subjects. She used a variety of nuts and seeds as reward and scaring with a broom as negative reinforcement. On the dish-row task "Grauhörnchen" was reported to have demonstrated a concept of fiveness and "Hans" one of sixness. Hassmann's simultaneous task required the differentiation between five lid-covered dishes each bearing a number from three to seven. These numbers were indicated by irregular dots that changed in size and position from trial to trial. One squirrel "Hexer" could differentiate the seven-lid from the three, four, five, and six-lids.

There are several observations by Hassmann which strongly suggest the presence of extraneous cues. A new fourness task was solved with an initial correctness equal to a previous threeness task involving 600 trials. Hassmann did not interpret this as a possible indicator of extraneous cues, but termed this behavior "a surprising success in learning." Odor was not controlled since negative dishes were not baited during a total of approximately 15,000 trials, in spite of the fact that Hassmann reported an aversion on the part of one of her animals from newly painted dishes (Hassmann, 1952, p. 299). On one occasion an unplanned odor control was reported. A squirrel pushed a positive lid aside, without "seeing" the peanut in it. It went to some negative dishes but returned later to the positive dish, opened it completely and obtained the bait. If "Peter" had smelled the peanut, Hassmann maintains, it would have continued to displace the positive lid on its initial attempt. Hassmann did not include peanuts in her previous list of rewards and it is difficult to determine the amount of acquaintance "Peter" had with this type of reward and its odor. The performance was very much like that which Tinklepaugh (1932) observed in monkeys when rewards were changed during a delayed reaction test.

Aside from odor the manual placing of all cups, lids, and sample dots could have presented additional extraneous cues. Hassmann's methodology should be scrutinized since she reported the successful learning of an oddity task in which the cue for solution was always numerosity. If this performance occurred without the aid of extraneous cues it would represent one of the highest conceptual achievements on subprimate level.

Wesley (1959) investigated numerosity in the rat in a successive task in which the animals were required to enter a "second" open alley without previously entering a "first" open one. The alleys, their location, and their total number changed from trial to trial. Some significant runs were obtained only by massing trials at the end of daily practice sessions, linked with nonreinforcement after an initial prolonged corrective training. Osgood[1] pointed out that it is possible the animal responded by avoiding the first open, negative door rather than by entering the second open one. Thus as in the case of Rossmann's fish the rats may have responded only to oneness.

The rats' capacity to discriminate by numerosity was further investigated by Wesley on a multiple serial visual discrimination apparatus. Rats were able to perform on a twoness task after approximately 100 trials and showed negative transfer to a subsequent threeness task. Discrimination of threeness was acquired but not maintained after the exclusion of triangularity.

Elephant

The visual learning capacity of an elephant was studied by Rensch and Altevogt (1953) who presented three- and four-dot patterns on stimulus cards. After almost 100 trials the elephant was able to distinguish correctly between irregular dots on a 3:4 discrimination problem, but only with constant arrangement of the

[1] C. E. Osgood, personal communication, October 1960.

stimulus dots. Since the positive three-dot pattern was always presented in one of seven arrangements and the negative four-dot pattern in one of five arrangements, the elephant could have solved the entire task by learning five different Gestalten, and it is therefore questionable whether the animal had the abstractive capacity the experimenters suggest.

Monkeys

Douglas and Whitty (1941) reviewed the literature of number appreciation in subhuman primates and tested four baboons in a visual discrimination experiment. They presented either one or two successive flashes and required a different response to each cue. When subsequently they were equated for duration the proportion of correct responses fell to a low value.

Kühn (1953) investigated the ability to differentiate visually between black dots of varied sizes and arrangements. His 2-year-old rhesus monkey "Lola" received a total of 18,718 trials within 439 working hours. Kühn used 50 discrimination cards per number throughout his experiment and presented these on training and on test trials. He reported learning to discriminate number on an 8:6 task, but it should be noted that the cards of the six series were presented 500 times prior to this task, always designating the negative stimulus. Responses may have occurred to individual cards and not necessarily by means of the number concept they presented. A similar type of learning may have been involved in the solution of the 8:7 task.

Hicks (1956) investigated the number concept in eight adolescent rhesus monkeys. His methodology was free of extraneous cues, since he introduced new and different stimulus cards during test trials. All of his animals performed above chance on a threeness problem, though some with rather moderate proficiency. Hicks compares his results with other studies and assumed that the 8:7 discrimination level observed by Kühn represents a true number concept but he had some doubt whether his own positive results indicated a number concept per se, since in all tests of number concepts the stimuli possess other characteristics than number. If, however, such a definition is employed no number concept per se could ever be demonstrated even on a human level, as stimulation always involves physical characteristics in addition to numbers. Heidbreder (1946), e.g., could not present twoness to her human subjects without involving objects, gestalten or size.

Conclusions

The performance of birds on a sevenness level has been compared to the human level of subitizing, an estimating of number without counting, where seven seems to form the average upper limit. (Jevons, 1871; Miller, 1956.) A re-examination of the methodology of these bird studies, however, makes such a comparison invalid and questions performance at any numerical level. Phylogenetically, the monkey would be expected to perform closer to the human level, but at present threeness is the only level unequivocally established with this species. The numerical capacity above threeness needs further investigation with monkeys, as numerosity in general needs to be studied further throughout the entire phylogenetic scale. To free future experiments from the influence of extraneous cues the presentation of the stimuli should be mechanical and should randomize time, distance, size, and amount of food and should control odor, noise, and other possible immediate cues. It is very likely that the use of rigid experimental controls will show that performance involving number concepts is congruent with the phylogeny of behavior as observed in other types of tasks.

REFERENCES

ARNDT, W. Abschliessende Versuche zur Frage des Zählvermögens der Haustaube. *Z. Tierpsychol.*, 1939, **3**, 88–142.

BRAUN, HILDEGARD. Über das Unterscheidungsvermögen unbenannter Anzahlen bei Papageien. *Z. Tierpsychol.*, 1952, **9**, 40–91.

DOUGLAS, J. W. B., & WHITTY, C. W. M. An investigation of number appreciation in some sub-human primates. *J. comp. physiol. Psychol.*, 1941, **31**, 129–142.

HASSMANN, M. Von Erlernen unbenannter Anzahlen bei Eichhörnchen. *Z. Tierpsychol.*, 1952, **9**, 294–321.

HEIDBREDER, E. The attainment of concepts: I. Terminology and methodology. *J. genet. Psychol.*, 1946, **35**, 173–189.

HICKS, L. H. An analysis of number concept formation in the rhesus monkey. *J. comp. physiol. Psychol.*, 1956, **49**, 212–218.

HONIGMAN, H. The number concept in animal psychology. *Biol. Rev.*, 1942, **17**, 315–337.

JEVONS, W. S. The power of numerical discrimination, *Nature, Lond.*, 1871, **3**, 281–283.

KOEHLER, O. "Zähl"-versuche an einem Kolkraben und Vergleichsversuche an Menschen. *Z. Tierpsychol.*, 1943, **5**, 575–712.

KOEHLER, O. The ability of birds to count. *Bull anim. Behav.*, 1951, No. 9, 41–45.

KÜHN, E. Simultanvergleich gesehener Mengen beim Rhesusaffen. *Z. Tierpsychol.*, 1953, **10**, 268–296.

LÖGLER, P. Versuche zur Frage des Zähl-Vermögens an einem Graupapagei und Vergleichsversuche an Menschen. *Z. Tierpsychol.*, 1959, **16**, 179–217.

MAROLD, E. Versuche an Wellensittichen zur Frage des "Zähl"-vermögens. *Z. Tierpsychol.*, 1939, **3**, 170–223.

MILLER, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.*, 1956, **63**, 81–97.

MORGAN, C. T. *Introduction to psychology.* New York: McGraw-Hill, 1956.

NEWMAN, J. R. *The world of mathematics.* Vol. 1. New York: Simon & Schuster, 1956.

RENSCH, B., & ALTEVOGT, R. Visuelles Lernvermögen eines indischen Elefanten. *Z. Tierpsychol.*, 1953, **10**, 119–134.

ROSSMANN, ANGELIA. Über das "Zähl"-Vermögen der Fische. *Z. Tierpsychol.*, 1959, **16**, 1–18.

SALMAN, D. H. Note on the number conception in animal psychology. *Brit. J. Psychol.*, 1943, **33**, 209–219.

SAUTER, ULRIKE. Versuche zur Frage des "Zähl"-Vermögens bei Elstern. *Z. Tierpsychol.*, 1952, **9**, 252–289.

SCHIEMANN, K. Vom Erlernen unbenannter Anzahlen bei Dohlen. *Z. Tierpsychol.*, 1939, **3**, 292–347.

THORPE, W. H. *Learning and instinct in animals.* Cambridge: Harvard Univer. Press, 1956.

TINKELPAUGH, D. L. Multiple delayed reaction with chimpanzees and monkeys. *J. comp. physiol. Psychol.*, 1932, **13**, 207–243.

WESLEY, F. Number concept formation in the rat. *Z. Tierpsychol.*, 1959, **16**, 605–627.

ZAHN, W. Über den Geruchssinn einiger Vögel. *Z. vergl. Physiol.*, 1933, **19**, 785–796.

(Received June 7, 1960)

Psychological Bulletin
1961, Vol. 58, No. 5, 429-440

HYPNOTIC AGE REGRESSION

AUBREY J. YATES[1]
University of New England, Australia

The alleged phenomenon of hypnotic age regression refers to the apparent fact that a subject (*S*) who is told under hypnosis that he is, e.g., 4 years old, may behave in a manner which is characteristic either of his own behavior at that age, or of children in general at that age. It should be noted carefully that the phenomenon may refer to the reactivation of behavioral characteristics of *S* himself; or may refer to a more general revivification of childlike behavior.

According to Platonow (1933), hypnotic age regression was first demonstrated clinically in 1893 by Kraft-Ebing. In spite of a good deal of clinical interest in the alleged phenomenon, experimental interest in the problem remained dormant until the publication of Platonow's report, in which he claimed that hypnotic age regression had been objectively demonstrated in three *S*s, using the Binet test, and that the general behavior of the *S*s in the regressed state showed characteristic childlike features. This conclusion was challenged by Young (1940), who claimed that *S*s were able in the waking state to reproduce the test and general behavior of a young child voluntarily, and with greater accuracy than hypnotized *S*s who had been regressed. Curiously, this challenging problem of the authenticity of hypnotic age regression has not received a great deal of attention since Sears (1943) first reviewed it briefly. A review of the literature suggests, however, that the problem is now much more clearly defined and a number of facts can be accepted as reasonably well-established. The principal controversies have centered around the disputes as to whether regression (partial or complete) can be demonstrated; if so, whether the regressed state can be simulated by *S* or represents a genuine reactivation of previous habit-systems or personality organizations; and finally, whether or not hypnosis is an essential part of the process.

This review will cover the types of measures which have been used to compare the test performance and behavior of the *S* in the waking state with that in the hypnotic state, the various conditions under which the performance is recorded, the principal established results, the main theories which attempt to account for the phenomenon, and the methodological problems involved. Some suggestions for future research will be made.

Types of Comparison

A distinction may be drawn between direct and indirect comparisons[2] of the waking and regressed states; and between the use of measures which are susceptible to simulation to a greater or less degree, and those which on the whole are not susceptible to simulation.[3]

[1] Now at the University of Western Australia. Acknowledgement is due to D. Howie for valuable criticisms of this paper.

[2] The problem is similar to that faced by the clinical psychologist attempting to measure deterioration (Yates, 1956).

[3] The distinction is, of course, an arbitrary one, but it does help in organizing the field.

Direct Comparisons

A few studies have directly compared *S*'s present regressed performance with his known performance on the same measure at a time corresponding to the regressed age level.

Simulable Measures. Sarbin (1950b) compared the performance of 12 *Ss* regressed to the age of 8 years with their performance on the same test when they were actually 8 years old. Similarly, Orne (1951) was able to compare the original and regressed drawings of one of his *Ss*.

Nonsimulable Measures. True (1949) regressed his *Ss* successively to ages 10, 7, and 4 years, and asked them what day of the week their birthday and Christmas Day fell on in that year. Best and Michaels (1954), and Reiff and Scheerer (1959) performed similar experiments, while the latter additionally attempted to reactivate factual information about childhood experiences (e.g., names of teachers and classmates) which *Ss* claimed to be unable to recall in the waking state.

These studies appear to represent the only sources of direct comparison thus far made.

Indirect Comparisons

In the majority of studies on hypnotic age regression the present regressed performance of *S* is compared with the average known performance of normal *Ss* of the age to which regression is induced.

Simulable Measures. There are three sources of evidence here. In the field of mental testing, investigators have used the Binet test (Keir, 1945; Platonow, 1933; Spiegel, Shor, & Fishman, 1945), the Wechsler-Bellevue Intelligence Scale (Kline, 1951), the Otis Performance tests (Kline, 1950), the Word Association Test (Dittborn, 1951; Keir, 1945), and various kinds of motor tasks, such as drawing a man, handwriting, etc. (Orne, 1951; Platonow, 1933). Second, the general behavior of *S* in the regressed state has been compared with that of normal children of that age-level (Keir, 1945; McCranie, Crasilneck, & Teter, 1955; Orne, 1951; Platonow, 1933). Third, *S* has been placed (in the regressed state) in situations known to evoke intense fear responses in many young children and his behavior observed (Kline, 1953a).

Nonsimulable Measures. Three kinds of information have been utilized. In the field of mental testing, the Bender Gestalt Test has been used (Crasilneck & Michael, 1957); and the Rorschach test (Bergman, Graham, & Leavitt, 1947; Keir, 1945; Mercer & Gibson, 1950; Norgarb, 1952; Orne, 1951), the argument, of course, being that it would be difficult for adults to simulate the performance of young children on these tests, especially in the case of the unstructured Rorschach test. Second, a number of physiological measures have been recorded while *S* was in the regressed state. These include the presence or absence of the plantar response (Gidro-Frank & Bowersbuch, 1948; True & Stephenson, 1951) and the Babinski reflex (McCranie et al., 1955); changes in indices such as blood pressure, pulse, and respiration rates, and psychogalvanic reflex (Kline, 1960; True & Stephenson, 1951); and changes in EEG characteristics (McCranie et al., 1955; True & Stephenson, 1951), the argument in the latter case being that there are characteristic differences between the records of adults and children.

Third, special mention must be made of the advances in technique recently reported by Reiff and Scheerer (1959). In accordance with Scheerer's general theoretical position, they abandoned the use of mental age tests, and utilized instead the notion of developmental levels, analyzing *S*'s *approach* to the solution of various problems, rather than his correct or incorrect *responses*. They used a number of ingenious tests which they consider it would be particularly difficult to simulate. Thus, in the Lollipops test, the regressed adult was given a lollipop after making mud-pies, while his hands were still dirty. If true regression had taken place, the child regressed to 4 years would, they argued, naturally not worry about his dirty hands when accepting the lolly; the adult simulating regression would do so. On the cognitive side, they used a Pledge of Allegiance test (writing the pledge after reciting it), a Clock test (telling the time), a Left and Right test (identifying left and right e.g., in persons sitting opposite), an Arithmetic test; Piaget's Hollow Tube Test (identifying the order in which colored beads will emerge from a hollow opaque tube after it has been rotated), and a Word Association test. On all of these tests, children show characteristic changes in modes of response with increasing age. On the Word Association test, for example, the most popular responses with children are quite different from those found in adults. They further argued that the adult who simulated a child's responses on this test (correctly or incorrectly) would show increased reaction time, since he would first need to inhibit his natural adult response tendency.

TESTING CONDITIONS

Six distinct testing conditions have been employed. Four of these may be regarded as control conditions for the two extreme conditions of performance in the normal waking state compared with performance in the suggested regressed state under hypnosis. The six conditions are:

1. Normal waking state—With some exceptions (e.g., Platonow, 1933; Sarbin, 1950b; Spiegel et al., 1945), nearly all investigators record performance in this condition.

2. Normal waking state, with deliberate (not simulated) attempted recall of earlier events—This control condition has seldom been used, though it is clearly essential to the validity of results such as those obtained in True's (1949) experiment.

3. Normal waking state, with instructions to simulate regression to a particular age-level—This control condition was used by Reiff and Scheerer (1959).

4. Standard hypnotic state—This represents a control for the effects of hypnosis per se, and has rarely been used.

5. Hypnotic state, with instructions to simulate regression to a particular age-level—This condition has been used only once experimentally (Crasilneck & Michael, 1957).

6. Hypnotic state, with direct suggestion by *E* that *S* is now a certain age.

No single study has used all of these conditions, and only one (Crasilneck & Michael, 1957) has used as many as four.

It would be an important requirement of any experiment in this field that the judgments of behavior in the various conditions should be made in ignorance of the particular condition in which *S* is placed at the time of assessment, i.e., the double-blind technique should be used.

Principal Results

Regression can be produced under hypnosis (Condition 6), the extent and accuracy of the regressed behavior being a matter of considerable dispute. Thus, Sarbin (1950b) using direct comparisons of Binet performances, found that under hypnosis not one of his nine hypnotizable *S*s achieved a mental age as low as that on the original test occasion. Using indirect comparisons, Crasilneck and Michael (1957) regressed their *S*s to the age of 4, but found that the Bender Gestalt drawings were rated by independent judges as comparable with those of 7-year-old children. It is not surprising, therefore, that inconclusive results have been generally reported for nonsimulable complex tests such as the Rorschach, except for obvious measures such as number of responses and form-quality (Bergman et al., 1947; Orne, 1951). On the other hand, it is important to notice that such regression as is achieved is often remarkably successful. Kline (1950) showed that in spite of a very significant decline in score on the Otis Performance tests under regression (from 59.2 in the waking state to 24.5 under regression to 8 years), the IQ at the regressed ages showed less variability than is normally found when the test is repeated on separate occasions.

Reiff and Scheerer (1959) reported almost uniformly perfect regression under hypnosis on all of the measures they used. Thus, on the Clock test, adults regressed to age 7 made errors characteristic of children of that age level, whereas the simulating controls did not make such errors.

Regression, however, does seem to become more accurate as the functions measured become more specific. True (1949), using 50 *S*s in the experiment described earlier, found that when regressed to ages 10, 7, and 4, 92%, 84%, and 62% of *S*s, respectively, correctly identified the day of their birthday; while 94%, 86%, and 87% of *S*s, respectively, correctly identified the day on which Christmas Day fell. Best and Michaels (1954) found negative results in a similar experiment, but they used only five *S*s and their procedure differed in important respects from that of True. Equally remarkable results were obtained by Reiff and Scheerer with the Word Association test. Thus, while adults uniformly respond to the word "man" with "women," children equally uniformly respond with the word "work." Their hypnotically regressed *S*s responded with the characteristic child's response while simulating *S*s continued to use the adult response word. McCranie et al. (1955) reported the reinstatement of the Babinski reflex in 3 of their 10 *S*s when regressed to the age of 1 month; while Gidro-Frank and Bowersbuch (1948) found significant changes in the plantar response, which were accompanied by changes in peripheral chronaxie. McCranie et al. (1955) did not, however, observe any significant change in EEG records in the regressed state.

Moody (1946), Ford and Yeager (1948), and Erickson (1937), have all reported the reinstatement of disabilities (wheal marks, homonymous hemianopsia, attacks of unconsciousness) under hypnotic regression; the disabilities no longer being present in the waking state.

The general behavior of the regressed *S* has been frequently reported as becoming more childlike, (e.g., Reiff & Scheerer, 1959) even to the extent of the appearance of the sucking response and loss of speech when regression is induced to a very early age (McCranie et al., 1955). In many respects too, the behavior is appropriate, not merely to the age to which *S* has been regressed, but in relation to the environment as it was at that time.

Regression can be produced in the waking state by asking *S* to simulate the suggested age (Condition 3). Under these circumstances, Crasilneck and Michael (1957) showed that Bender Gestalt drawings will reflect the simulated age level less successfully than is the case with hypnotic regression; and this finding has been amply substantiated by Reiff and Scheerer (1959).

Regression can be produced in the hypnotic state by asking *S* to simulate the suggested age (Condition 5). Under this condition, Crasilneck and Michael (1957) showed that Bender Gestalt drawings will not be significantly different in quality from those produced by direct suggestion under hypnosis.

The results obtained by Crasilneck and Michael (1957) for the waking state, waking state with simulated regression, hypnotic state with simulated regression, and hypnotic state with induced regression, for the Bender Gestalt test indicated that regression was not complete under any of the conditions, that regression could be simulated, and that hypnosis facilitated the production of regressed behavior. Reiff and Scheerer (1959) did, however, find complete regression to appropriate developmental levels.

There is some indication that emotional regression can be induced. Kline (1953a) regressed a female *S* to the age of 3 years and placed her in situations frequently found to produce intense fear responses in young children (being left alone, entering a dark passage, seeing a strange person oddly dressed, sudden appearance of a live snake, sight of a headless doll, and presence of a live mouse). All but the first and last situations produced realistic fear reactions in the regressed, but not in the waking, state, including involuntary urination. Reiff and Scheerer (1959), analogously in a play situation, found less repugnance to eating with filthy hands under regression.

The mere induction of hypnosis itself does not produce regressed behavior in the normal *S* (Bergman et al., 1947; Kline, 1953a), though there are clinical reports of "spontaneous regression" under hypnosis (Gill, 1948; Keir, 1945; Schneck, 1955).

Sarbin (1950b) has reported a correlation of +0.91 between a regression index and degree of hypnotizability. His hypnotizable *S*s were regressed under hypnosis, and later asked to simulate regression in the waking state. A regression index (RI) was computed for *S* according to the formula:

$$\mathrm{RI}=\left[\frac{\text{MA (simulated regression)}}{\text{MA (original test)}}-\frac{\text{MA (hypnotic regression)}}{\text{MA (original test)}}\right]\times 100$$

We may conclude from this brief survey of results that a prima facie case appears to have been made out for the assertion that under some conditions certain adults behave in ways which are characteristically those of children, although many of the details remain to be filled in.

Theories of Hypnotic Age Regression

Three theories have been proposed in attempts to account for the above results.

Neurological Theories

Platonow (1933) explained regression in terms of what he called Pavlov's "true physiology of the brain," using especially the notion of words as conditioned stimuli producing physiological, biological, and psychological changes:

> the suggestion of previous ages brings forth a real organic reproduction of the engrams, the formation of which belongs to the earlier periods of the individual's life (p. 205).

Regression is facilitated under hypnosis because the latter involves general inhibition of the cortex except for the area receiving auditory impulses. Under these conditions, the auditory stimulus (suggestion of regression) most readily activates the appropriate engrams. This theory would appear to be derived from the much older distinction between cortical and subcortical brain-processes, the latter mediating primitive responses. It is interesting to note that McCranie et al. (1955) assert that lesions of Brodman's Area 4 result in the restoration of the Babinski reflex in chimpanzees and man, while simultaneous bilateral ablation of Areas 4 and 6 produces infantile motor behavior in lower primates.

Kline (1953b, 1954) has proposed a "neuropsychological" theory, derived from the experimental observation that what he terms habit progression, as well as habit regression, can be demonstrated (Kline, 1951; Rubenstein & Newman, 1954). Thus, in one study by Kline (1951), a 22-year-old woman was able, under hypnosis, to produce the typical Wechsler-Bellevue record of a 65-year-old woman, even to the extent of obtaining a characteristic Deterioration Index score for that age-group. Kline's theory (1953b) postulates that

> the actual state involved in such activity is not regression, not progression, but a central state of perceptual release or disorientation which permits activity in any dimension or direction of time-space orientation (p. 26).

Under hypnosis, there occur what Kline (1953b) calls "directional alterations from a central process" (p. 25) and he lays particular stress on the importance of transference relationships between *S* and the hypnotist. He does not, however, regard the phenomenon as simply involving role-taking (see below) and his theory, in particular does not regard the evidence obtained from psychometric measures as crucial, though he does not deny their relevance. As at present formulated, Kline's theory would seem to be too general to be amenable to disproof.

Habit-Reactivation

Hypnotic age regression may be regarded as a special instance of instrumental act regression. In the latter, if *S* possesses alternative response patterns to a given stimulus, the stronger will normally occur. If, however, the stronger is prevented from occurring, then the inhibited response pattern will be reactivated. It is possible that, in some way as yet unknown, hypnotic suggestion of regression may inhibit current response patterns, and hence permit the reactivation of "forgotten" response patterns. Although this theory as presented here is very general, it is surprising that no consideration has yet been given by any worker in this field to the relationship between hypnotic age regression and instrumental act regression. Contrariwise, it would be predicted that newly acquired contioned responses would be lost in hypnotic age regression. McCranie and Crasilneck (1955) attempted to test this hypothesis by setting up voluntary hand withdrawal, and involuntary eyeblink, conditioned responses. Under hypnotic regression, the former disappeared, the latter did not. Similar results were reported by LeCron (1952).

Reiff and Scheerer (1959) have put forward a theory of hypnotic age regression which is derived from a general theory of remembering, but which could equally be regarded as a more general theory of habit reactivation than the instrumental act regression theory. According to them:

> the act of recall becomes also an act of contemporary reconstruction of the past event, to a large extent dependent upon the state of the person at the time of the recall (p. 15).

Such memories can take the form either of remembrances or of memoria. In the waking state, remembrances are memories with a conscious autobiographic index (i.e., experienced as "being in my past") and involve therefore personal continuity. Memoria are memories without a current autobiographic past reference (e.g., motor skills, vocabulary, etc.). Although the distinction is not absolute, remembrances are usually related to an experiential context, whereas memoria are related to an environmental context. Both kinds of remembering may arise voluntarily or involuntarily.

Hypnotic age regression

> makes possible a reinstatement of the forgotten personal past either in the form of remembrances, or in the form of memoria and earlier ego apparatuses (p. 52). . . . [In general] since . . . memoria are without an experiential auto-biographic index, the ego can more easily activate appropriate memoria than remembrances (p. 49) . . . . [but] the age-regressed subject may remember events with the experience of an auto-biographic index. However, here the reference point is no longer the *actual* present but that point in the autobiographic past to which the subject has been regressed (p. 52).

Remembrance or memoria reactivated in this way are always involuntary.

In hypnotic age regression, therefore, the attempt is made to reactivate the general environmental and experiential contexts of the *S* at the age regressed to. If these remembrances can be activated sufficiently strongly, then individual items of behavior may be reactivated in the form of memoria. It should be noted that in regression, memoria and remembrances are experienced as occurring here and now, whereas in the normal state they can both be referred to the past, though only remembrances will have autobiographical references.

Role-Playing

This viewpoint has been well expressed by the assertion that

> we may formulate the concept of age regression by saying that the prevailing psychological condition enables the individual to take the role appropriate to the imagined world (Orne, 1951, p. 220).

The most important exponent of this theory is Sarbin (1950a; Sarbin & Farberow, 1952). Sarbin's theory holds that all human interaction involves both *S* and *E* indulging in role-playing. The validity of role-playing depends upon a number of factors, including the validity of *S*'s perception of the interaction situation, the aptitude of *S* for playing a particular role, and the current organization of the self. Thus, if age regression is to appear

> the *S*'s perception of the child's role must have some veridical properties; there must be present evidence of the role-taking aptitude, and the assigned role must not be incongruent with the *S*'s current self-perceptions. (Sarbin & Farberow, 1952, p. 119.)

Sarbin's theory has resulted in a number of interesting predictions, for example, that an *S* whose self-organization is relatively undeveloped should show greater regression than an *S* whose self-organization is normal. Unfortunately, little concrete evidence in support of Sarbin's theory, in so far as it relates to age regression, has yet been produced.

Methodological Problems

Many methodological problems present themselves in connection with hypnotic age regression. They may be grouped into five areas.

Types of Control

Mention has already been made of some of the conditions under which testing has been carried out. Logically, at least three sets of factors can be varied, giving eight possible combinations of testing conditions. Thus, *S* may be tested in the hypnotic or waking state, the comparisons made may be direct or indirect, and regression or simulation may be attempted. Additionally, account must be taken of the effects on performance of hypnosis, and the waking state, per se.

Criteria for Regression

It is obvious that complete biological regression is impossible (the *S* in the regressed state does not, for example, diminish in stature). The critical question therefore becomes: "Does the hypnotically regressed adult perform as he *imagines* a child to function or is his regressive behavior a revival of memoria, i.e. of certain aspects of his previous functioning?" (Reiff & Scheerer, 1959, p. 83). It should be realized at the outset that the fact that hypnotic age regression may be incomplete does not in itself prove the validity of the role-playing theory; nor does the demonstration of successful role-playing in itself disprove the validity of regressive phenomena. The most satisfactory evidence for the validity of regression would involve the demonstration that *S* performed in the regressed state in a manner similar to his behavior at that age when a child (thus involving *direct* comparison) together with the demonstration that he was unable, as an adult, either under hypnosis or in the waking state, to simulate the appropriate behavior. Evidence of this kind has thus far been presented only in isolated cases. It is not here intended to deny, of course, that in most instances of hypnotic age regression, both true regression and role-playing may be simultaneously involved.

Hypnotic Technique

Several important points have commonly been neglected. Criteria for measuring the depth of trance are usually not reported in sufficient detail. The speed with which regression is induced is possibly a critical variable and may explain the failure of Best and Michaels (1954) to repeat the results of True (1949). It is probably important to reinstate the earlier period gradually, rather than suddenly. The role of the hypnotist is often neglected—it has been argued that true regression becomes more likely if the hypnotist transforms himself into some person familiar to *S* at the regressed age, or at least into a neutral figure. Reiff and Scheerer (1959) lay particular stress on the importance of the instructions given to *S*, who should be regressed to a specific date (e.g., a birthday) and not merely to a particular year. Fluctuations in performance should be controlled as far as possible by instructing *S* not to deviate from his regressed age-level.

Selection of Ss

Very little attention has been paid to this important aspect of the problem. Reiff and Scheerer (1959) lay great stress on the difficulty of obtaining suitable *S*s who must be relatively free from anxiety (severe anxiety about events happening at age 4, for example, might well lead to resistance to regression to that age), must be suitably motivated, and, of course, satisfactorily hypnotizable. Such *S*s are relatively rare. Selection of control *S*s has been even more neglected. Reiff and Scheerer, for example, give few details about their control *S*s and do not seem to realize that the experimental and control *S*s should have been carefully matched on all relevant variables (including hypnotizability).[4] This failure was especially serious in that, with the exception of two measures, Reiff and Scheerer did not control for performance in the waking state, apparently assuming that all their *S*s would perform normally.

Selection of Tests and Measures

The search for nonsimulable tests has been markedly improved by the suggestions of Reiff and Scheerer who, in addition to the tests they used, have made a number of ingenious suggestions for further research. The most significant measures thus far utilized are undoubtedly the Birthdays test of True (1949) and the Word Association test used by Reiff and Scheerer (1954). The latter, however, prefer the use of developmental schedules to mental age scales, and are interested more in the process of solution than in the solution itself. While the distinction is not entirely academic, it certainly has not the importance attributed to it by Reiff and Scheerer. Two minor points may be noted: the tasks, measures, or developmental schedules used, should be appropriate to the age level regressed to; and care should be taken to prevent *S* from giving no response ("I don't know") except where such a response is explicitly predicted.

It may safely be said that no fully adequate experiment has been carried out in this field. Thus, the most recent study by Reiff and Scheerer (1959), although admirable in many respects, contained a number of serious faults: lack of control for performance in the waking state; failure to match experimental and control groups; and repeated testing at different age levels of the same *S*s in the experimental, but not in the control group. Even more serious, it is clear from the description given of the experimental procedure, that the authors were aware, in the testing situation, of which *S*s had been hypnotized, and which had not.

Discussion

The potential importance of the phenomenon of hypnotic age regression can scarcely be overestimated. Apart altogether from its possible value in general psychotherapy (Kline, 1950), and its usefulness in particular for the treatment of war neuroses by regressing the patient to the traumatic situation and making him relive the experience, it does not seem to have been generally realized that the technique itself could

[4] Their five hypnotized *S*s were chosen from an original group of over 100 *S*s.

provide a crucial test of the theory that learned responses are never "destroyed," but only supplanted and remain available for activation under appropriate circumstances. In light of this, it is surprising how little experimental work has been carried out in this area. Furthermore, a good deal of this work can hardly be said to attain even minimally acceptable levels of methodological adequacy.

Any acceptable theory of hypnotic age regression must take account of the apparent facts that regression can be simulated in the waking state; that the amount of regression is similar whether it is simulated under hypnosis, *or* suggested under hypnosis; and that regression becomes more "accurate" as the response regressed to becomes more specific. It seems likely that neither the role-taking, nor the habit reactivation theories, taken separately, will account satisfactorily for the observed facts. Thus, the role-taking theory is clearly embarrassed by findings such as those of True (1949) in relation to the recall of factual information under hypnosis, those of McCranie et al. (1955) in relation to the reactivation of physiological responses, and those of Reiff and Scheerer (1959) in relation to the Word Association test. It is probably necessary to recognize that both theories must be invoked, each accounting for some, but not all, of the facts. In this connection, it may be noted that a satisfactory explanation of the facts awaits the formulation of a valid *general* theory of behavior under hypnosis. Since, however, workers in this field are still struggling to elucidate basic concepts (Barber, 1958; Sutcliffe, 1960), it is probable that a more careful and thorough examination of the *phenomena* encountered in hypnotic age regression will provide data highly relevant to this aim.

We may conclude, therefore, with some general suggestions concerning future research in this field. First, a crucial area of research is the problem of partial versus complete age regression. It seems clear that complete regression would be extremely unlikely in relation to complex items of behavior, since early habit-structures involving complex skills would surely be affected by subsequent growth of the skill, if only through the process of retroactive inhibition. On the other hand, relatively isolated items of knowledge (such as knowing on what day one's fourth birthday fell) might easily survive relatively unchanged by subsequent learning, to be reactivated under appropriate conditions. Mention has already been made of the necessity for a close examination of the conditions under which regression is induced.

Second, more attention should be paid to an analysis of simple aspects of behavior, rather than complex ones. For example, instead of using measures such as the Binet, attention could be concentrated on, e.g., developmental schedules, which objectively record the presence or absence of specific items of behavior at different age levels. The suggestions for research made by Reiff and Scheerer (1959) are particularly valuable in this connection. The use of conditioning techniques, as exemplified by the study of McCranie and Crasilneck (1955) should also yield crucial information.

Third, much more attention should be paid to a careful description of the total behavior of *S* in the regressed state. Much has been made of the fact that *S* behaves in a manner appropriate to his regressed age. Almost invariably, however, the description is highly selective. For example, as Orne (1951) has pointed out, regression implies also that all knowledge acquired subsequently to the age to which *S* has been regressed should be unavailable. In other words, *S* should no longer be cognizant of current affairs, political, social, or otherwise. It is extremely curious that no information of a concrete nature on this vital point is available, except for a few vague, general assertions.

Fourth, no attention has been paid to the study of the behavior of *S* in the regressed state over a substantial period of time. Practically all investigators have restricted their observations to laboratory situations.

Fifth, the fact that hypnosis apparently facilitates simulated regression, but the addition of direct suggestion does not produce an improvement over hypnotically simulated regression requires further exploration. Thus far, evidence relating to this important point is restricted to results from a single study (Crasilneck & Michael, 1957).

The importance of the phenomena encountered in hypnotic age regression, and the advances in technique which characterize the investigations of Reiff and Scheerer (1959) should surely lead to a revival of interest in this problem.

REFERENCES

BARBER, T. X. The concept of hypnosis. *J. Psychol.*, 1958, **45**, 115–131.

BERGMAN, M. S., GRAHAM, H., & LEAVITT, H. C. Rorschach exploration of consecutive hypnotic chronological age level regressions. *Psychosom. Med.*, 1947, **9**, 20–28.

BEST, H. L., & MICHAELS, R. M. Living out "future" experience under hypnosis. *Science*, 1954, **120**, 1077.

CRASILNECK, H. B., & MICHAEL, C. M. Performance on the Bender under hypnotic age regression. *J. abnorm. soc. Psychol.*, 1957, **54**, 319–322.

DITTBORN, J. Words associated to different age levels suggested under hypnosis. *Rev. Psiquiat., Santiago*, 1951, **16**, 105–107.

ERICKSON, M. H. Development of apparent unconsciousness during hypnotic reliving of a traumatic experience. *Arch. Neurol. Psychiat., Chicago*, 1937, **38**, 1282–1288.

FORD, L. F., & YEAGER, C. L. Changes in the electroencephalogram in subjects under hypnosis. *Dis. nerv. Syst.*, 1948, **9**, 190–192.

GIDRO-FRANK, L., & BOWERSBUCH, M. K. A study of the plantar response in hypnotic age regression. *J. nerv. ment. Dis.*, 1948, **107**, 443–458.

GILL, M. Spontaneous regression on the induction of hypnosis. *Bull. Menninger Clin.*, 1948, **12**, 41–48.

KEIR, G. An experiment in mental testing under hypnosis. *J. ment. Sci.*, 1945, **91**, 346–352.

KLINE, M. V. Hypnotic age regression and intelligence. *J. genet. Psychol.*, 1950, **77**, 129–132.

KLINE, M. V. Hypnosis and age progression: A case report. *J. genet. Psychol.*, 1951, **78**, 195–206.

KLINE, M. V. Childhood fears in relation to hypnotic age regression: A case report. *J. genet. Psychol.*, 1953, **82**, 137–142. (a)

KLINE, M. V. Hypnotic retrogression: A neuropsychological theory of age regression and progression. *J. clin. exp. Hypnosis*, 1953, **1**, 21–28. (b)

KLINE, M. V. Living out "future" experience under hypnosis. *Science*, 1954, **120**, 1076–1077.

KLINE, M. V. Hypnotic age regression and psychotherapy: Clinical and theoretical observations. *Int. J. clin. exp. Hypnosis*, 1960, **8**, 17–42.

LECRON, L. M. The loss during hypnotic age regression of an established conditioned reflex. *Psychiat. Quar.*, 1952, **26**, 657–662.

MCCRANIE, E. J., & CRASILNECK, H. B. The conditioned reflex in hypnotic age regression. *J. clin. exp. Psychopathol.*, 1955, **16**, 120–123.

MCCRANIE, E. J., CRASILNECK, H. B., &

TETER, H. P. The electroencephalogram in hypnotic age regression. *Psychiat. Quart.*, 1955, **29**, 85–88.

MERCER, M., & GIBSON, R. W. Rorschach content in hypnosis: Chronological age level regression. *J. clin. Psychol.*, 1950, **6**, 352–358.

MOODY, R. L. Bodily changes during abreaction. *Lancet*, 1946, **2**, 934–935.

NORGARB, B. A. Rorschach psychodiagnosis in hypnotic regression. In L. M. LeCorn (Ed.), *Experimental hypnosis*. New York: Macmillan, 1952. Pp. 178–214.

ORNE, M. T. The mechanisms of hypnotic age regression. *J. abnorm. soc. Psychol.*, 1951, **46**, 213–225.

PLATONOW, K. I. On the objective proof of the experimental personality age regression. *J. gen. Psychol.*, 1933, **9**, 190–209.

REIFF, R., & SCHEERER, M. *Memory and hypnotic age regression*. New York: International Univer. Press, 1959.

RUBENSTEIN, R., & NEWMAN, R. The living out of "future" experiences under hypnosis. *Science*, 1954, **119**, 472–473.

SARBIN, T. R. Contributions to role-taking theory: I. Hypnotic behavior. *Psychol. Rev.*, 1950, **57**, 255–270. (a)

SARBIN, T. R. Mental age changes in experimental regression. *J. Pers.*, 1950, **19**, 221–228. (b)

SARBIN, T. R., & FARBEROW, N. L. Contributions to role-taking theory: A clinical study of self and role. *J. abnorm. soc. Psychol.*, 1952, **47**, 117–125.

SCHNECK, J. M. Spontaneous regression to an infant age level during self-hypnosis. *J. genet. Psychol.*, 1955, **86**, 183–185.

SEARS, R. R. Survey of objective studies of psychoanalytic concepts. *Soc. Sci. Res. Coun. Bull.*, 1943, No. 51.

SPIEGEL, H., SHOR, J., & FISHMAN, S. An hypnotic ablation technique for the study of personality development. *Psychosom. Med.*, 1945, **7**, 273–278.

SUTCLIFFE, J. P. "Credulous" and "sceptical" view of hypnotic phenomena. *Int. J. clin. exp. Hypnosis*, 1960, **8**, 73–102.

TRUE, R. M. Experimental control in hypnotic age regression states. *Science*, 1949, **110**, 583–584.

TRUE, R. M., & STEPHENSON, C. W. Controlled experiments correlating electroencephalogram, pulse, and plantar reflexes with hypnotic age regression and induced emotional states. *Personality*, 1951, **1**, 252–263.

YATES, A. J. The use of vocabulary in the measurement of intellectual deterioration: A review. *J. ment. Sci.*, 1956, **102**, 409–440.

YOUNG, P. C. Hypnotic regression: Fact or artifact? *J. abnorm. soc. Psychol.*, 1940, **35**, 273–278.

(Received August 2, 1960)

VOL. 58, NO. 6 NOVEMBER 1961

Psychological Bulletin

POPULATION DENSITY AND ENDOCRINE FUNCTION[1]

D. D. THIESSEN
University of California, Berkeley
AND DAVID A. RODGERS[2]
Scripps Clinic and Research Foundation

Early evidence (Crew & Mirskaia, 1931) suggested that the population size of many mammalian species and especially of rodents is self-limiting. In 1952, Calhoun demonstrated density limitation in a confined population of Norway rats. The population he observed never exceeded 200, even though he estimated the growth potential in terms of shelter, space, and food to be well over 5,000. Subsequently, a number of studies have demonstrated that the reproductive capabilities of rodents living in high-density populations are impaired (e.g., Chitty, 1955; Christian, 1959c; Christian & LeMunyan, 1958; Hoffman, 1958; Kalela, 1957; Louch, 1956; Southwick, 1955a, 1955b; Strecker & Emlen, 1953). Research into the mechanisms by which density limitation is accomplished reveals an interaction between density, endocrine function, and behavior that has major implications for the behavior theorist working with animal subjects. This research is reviewed in the present paper.

In a rather comprehensive theory based on Selye's conception of a general adaptation syndrome, Christian (1950) implicated the endocrine system in limitation of population density. He proposed that the observed triphasic population cycle consisting of an initial growth of population followed by a period of stability and then a period of decline could be accounted for by a stimulus feedback reaction described by Selye (1946), involving the endocrine system and particularly a pituitary-adrenocortical-gonadal axis. According to Selye, certain pituitary-adrenal-gonadal effects are produced by all general stressors and are in proportion to the severity of the stress. These effects consist in part of hyperactivity of the pituitary and adrenals and hypoactivity of the gonads. Christian reasoned that if population density were a stressor, it would be inversely related to gonadal activity and therefore to reproductive behavior, as well as to other factors affecting survival. Such relationships could account for the apparently self-limiting nature of density of population and would help to account for the triphasic population cycle. Under conditions of low density of population and in otherwise favorable circumstances, gonadal and reproductive activity would be high, resulting in an expanding population. The increasing population density, acting

[1] This investigation was supported in part by a fellowship MF-11,174 from the National Institute of Mental Health, United States Public Health Service.

[2] Formerly University of California, Berkeley.

as an increasing stressor, would eventually reduce reproduction to the point that deaths would match births. The population would reach equilibrium at that point and would enter the second, stable phase of the population cycle. Such stability would be maintained until the population was subjected to an additional stressor, such as increased daylight or increased cold occurring with seasonal change. The additional stressor could destroy the equilibrium and precipitate a more or less rapid decline of population, partially by its effects on reproduction rate and partially by other lethal effects of the increased stress.

Natural Populations

Support for Christian's theory was provided by the discovery of a relationship between stages of the density cycle and adrenal weight in natural populations of Norway rats (Christian & Davis, 1956). Rats from 21 Baltimore city blocks were systematically sampled and their population numbers estimated over a period of months. Since the rats seldom cross streets, each block was essentially an independently varying population. At time of sampling, each of these populations was classified as belonging to one of five successive stages of a population cycle: low stationary, low increasing, high increasing, high stationary, and decreasing. Beginning with the low increasing stage, a progressive increase in adrenal size was found for the successive stages. The relationships in the low stationary stage were somewhat ambiguous, but perhaps appropriately so if it is remembered that this stage constitutes the end of the population cycle as well as its beginning. To the extent that adrenal size correlates with adrenal activity, the results strongly suggest a progressive increase in stress, and in endocrine response to it, as the population cycle progresses. However, in this study, significant weight changes were not found in the thymus and pituitary glands, which normally show response to prolonged stress. Since nutritive elements were found in abundant amounts, social rather than strictly biological factors were presumed to be primary in determining the differences in adrenal weights. In a study of a rural population of Norway rats, Christian and Davis (Christian, 1959c) found a correlation of .90 between population density change and adrenal weight change. In this study, pituitary weight also changed with population size and correlated .99 with adrenal weight.

Louch (1956) reported similar findings in two natural populations of meadow voles. Population densities were estimated from monthly live trappings. In both populations, adrenal weight correlated positively with density. Eosinophil count, a blood-fraction measure known to vary inversely with adrenocortical activity, was found to have an inverse relationship to population density. The absence of apparent food shortages again suggested that nonnutritive factors were primarily responsible for the observed correlations.

Population Size

Christian has suggested that a logarithmic relationship exists between size of population and the endocrine or related effects. A number of his laboratory and field researches support this contention. In one study (Christian, 1955a), he placed weanling male mice in groups of 1, 4, 6, 8, 16, and 32 for one week. Adrenal weight in all cases except for the population of 32 showed a linear relationship to the logarithm of the population size. Adrenal weight for the largest population showed a decline from the next largest. This decline was initially interpreted as due to "social structure deterioration," representing some decrease in stress at the greatest density. More recently, however, Christian (1959c) has found that the relative decrease in adrenal weight at this high density is due to a loss in lipid content of the cortical cells, indicating an intense activation of the adrenocortex. The trend of increased adrenocortical activation with increased density would therefore appear to hold for all limits tested. Wild and tame mice show similar types of response to artificially established densities (Christian, 1955b), although both initial adrenal size and response to density is greater in the wild mice. Both responded to increased density by increased adrenal size, sex gland atrophy, and thymus atrophy (another index of stress).

Christian (1956a) found similar relationships in a study of free-growing populations of laboratory reared mice. Beginning with a few pairs in large cages amply supplied with food and water, some populations were allowed to reach an apparently self-determined asymptote. This was much below the number of animals that could be supported by the food and cover available. Other populations were allowed to reach approximately half of the expected maximal density of population. Still other animals were derived from segregated pairs and were maintained as segregated pairs following weaning. In the free-growing populations, the growth curves were sigmoid. Birth rate and survival of infants declined in proportion to the logarithm of the size of the population. As was found in both the wild and the artificially constituted populations, increased adrenal weight was found to be associated with increased density of population. Histological examination revealed that the increase in weight was due primarily to hypertrophy and hyperplasia of the zona fasciculata, suggesting greater adrenocortical functioning. In the young male mice, part of the higher adrenal weight was due to delayed involution of the X zone, a transitory layer of the adrenocortex, found in young mice. Since this involution is brought about by androgens, the observed delay in involution suggests that androgen production, and therefore the onset of puberty, occurs at a later age in male mice from high-density populations. It further supports the assumption that a pituitary-adrenal-gonadal interaction system is involved in control of population density. Reproductive organs of the mature high-population-density males were also lighter and spermatogenesis was partially suppressed as compared with the low-density controls.

Other results implicating endocrine involvement in population control were also found. The decrease in survival rate of infants in the dense populations was attributed to deficient lactation of the mothers. The infants who died usually did so in 10 to 14 days after birth and were found to be uninjured but with empty stomachs. The survivors were weaned early, appeared grossly stunted, and were in poor condition. If production of prolactin, one of the gonadotrophic hormones of the pituitary, is suppressed, along with suppression of the other gonadotrophic hormones, then reduced lactation and the observed infant mortality and stunting would be expected. Suppression of the gonadotrophins would also account for the observed reduction in numbers of pregnancies and numbers of embryos per pregnancy in the denser populations and for the increase in numbers of resorbing embryos per pregnancy observed at autopsy. These suggestions are in general agreement with the recent findings of Helmreich (1960). In this case grouped female deer mice showed increased resorption of implanted embryos, although the incidence of pregnancy and the number of embryos implanted were not different from those of the isolated controls.

Of considerable interest is Christian's additional finding, consistent with Chitty's (1955) speculations, that the effects of decreased body weight, intrauterine mortality, decreased litter size, and reduced ability to lactate were still observable in the first and second generation offspring of grouped animals. The effects presumably were transmitted as a function of decreased, or nutritionally altered, milk supply of the mother.

Louch (1956) carried out a study of three freely growing but confined populations of meadow voles that in many respects parallels Christian's studies of house mice. During the period of observation, the three meadow vole populations had access to abundant food and nesting supplies. The growth curves were sigmoid, much like those found by Christian, although the rate of growth varied considerably among populations. Although number and size of litters were not significantly correlated with density, several other factors that tended to limit population size did vary as expected. Litter mortality was high under conditions of dense population. This was attributed to the mother's reduced ability to lactate, to her greater tendency to eat or abandon her pups, and to increased trampling and disturbance of the litters by other animals. Adult mortality also correlated positively with size of population and was most pronounced during periods of population decline. The correlation was due at least in part to an apparent increased susceptibility to disease in dense populations. Amount of fighting and wounding increased with density. There was a tendency for fecundity, as measured by number of mice with scrotal testes or perforate vaginae, to correlate inversely with density. This inverse correlation was significant in one population, approached significance in another, and was opposite in sign and insignificant in the third. At high densities, males competed aggressively for females in heat, by chasing, fighting, and pushing each other away from the female. As a result, the number of mountings increased but few mountings led to completed copulation. Lowered eosinophil counts at the higher densities, together with the other findings, suggested, as do Christian's results, the direct involvement of the pituitary-adrenal-gonadal axis in the dynamics of population density.

Although many effects appear to follow a logarithmic relationship to population size, factors other than density alone can be important. Southwick (1955a) reports, for example, that different freely-growing populations of wild trapped house mice confined under essentially similar conditions varied as much as five fold in maximum size of population attained. He attributed the differences to uncontrolled social and genetic factors. Christian's finding (1955b) that wild mice show a more marked adrenal response to density than do laboratory mice suggests the importance of genetic factors. Other investigators, while not questioning the fact of density limitation, have questioned whether or not size of population is the crucial variable. Several studies have been directed toward assessing possible alternative explanations.

Wounding and Social Rank

In mice living 4, 8, or 16 to a cage, Southwick and Bland (1959) found no significant differences among the groups in adrenal weight unless wounded animals were compared with nonwounded. The wounded animals were found to have significantly heavier adrenals. They conclude that wounding is the essential operant in adrenocortical change and that higher density acts indirectly to increase adrenal size by creating a situation in which fighting and wounding is more likely to occur. Chitty, Chitty, Leslie, and Scott (1956) found similar evidence. Young male voles were put in contact with old mated pairs for periods of about 2 hours a day for several days. Fighting, chasing, and wounding typically occurred. The more severely wounded animals had a higher liver, spleen, and adrenal weight and a smaller thymus and body weight than did the less severely wounded. Clarke (1953), too, found similar effects when voles were introduced into cages containing a pair of "resident" animals. The newly introduced voles were viciously attacked and wounded; the longer the period of exposure, the more severe were their wounds and glandular changes. These studies implicate the physical effects of fighting and wounding as crucially important.

Contradictory results have been obtained in other studies, in which no relationship was found between amount of wounding and adrenal size and in which glandular effects occurred with little or no fighting and no wounding at all. Christian (1959b) measured adrenal hypertrophy and presence or absence of scarring in 50 populations of four, five, or six albino male mice each. When adrenal weight was corrected for body weight, adrenal hypertrophy, found after 1 week of grouping, did not reflect either the severity of fighting or the amount of injury received. Barnett (1955), working with two strains of rats, took movies of fighting behavior, territoriality establishment, and the working out of hierarchies within groups. Histological examination of the adrenals revealed hypertrophy in the subordinate animals only; this hypertrophy was related to the social position within the group but not to the amount of fighting. Christian and Davis (Christian, 1959c) also found that dominant Norway rats showed little adrenal hypertrophy, even though they fought as much as or more than subordinate animals that did show adrenal change. Southwick and Bland (1959) found adrenal hypertrophy more likely to occur in males housed with females than in males housed with other males, even though fighting was not observed and wounding did not occur in either case. When wild male house mice were grouped six to a cage 4 hours a day for several days (Davis & Christian, 1957), a significant negative relationship was found between social rank and adrenal weight. A similar relationship was found by Vandenbergh (1960) using eosinophil levels and adrenal weights as indices of adrenocortical activity.

These studies suggest that social rank is an important factor in determining endocrine response. Amount of wounding and social rank tend to be related in recently grouped populations, so that some correlation with wounding might be expected under some circumstances. Also, the height of the pyramid towering over a low-status animal, as well as the number of low-status animals, is a factor of the size of the population and would lead to an expected correlation between average endocrine response and population size. If social status were crucially important in determining the endocrine response, then the reproductive capacity and stress vulnerability of the low-status animals would be affected first in an expanding population. A selective advantage would therefore accrue to those characteristics making for high status in the population.

An interesting parallel appears between these data and data showing that hormonal variations in the blood stream are related to changes in dominance. If the initial status differences among animals are not too wide or too greatly solidified by learning, the administration of androgen to low-ranking normal and castrated animals increases the dominance status in both the male and the female (see Beach, 1948; Bindra, 1959).

Living Space

In studies of white Leghorn chickens, Siegel (1959a, 1959b, 1960) placed different numbers of birds in equal-sized pens and found that the more crowded groups had larger adrenals and produced fewer eggs. In some comparisons, he found smaller pituitary weights and histochemical evidence of greater adrenocortical secretion in the more crowded groups. Siegel ascribes these relationships to differences in floor space per animal. They are highly consistent with the data from rodent populations in relating density to endocrine and reproductive response. In Siegel's studies, as in many of the rodent studies, population size and living space per animal are confounded, leaving open the possibility that size of population rather than living space is the crucial variable affecting endocrine response. Other studies tend to rule out the importance of living space as an independent variable. Christian found that the same positive relationship between population size and adrenal weight held for mice even when floor space was increased 42 times. In an unpublished study, we have found that the relationship held when living space per mouse was exactly equated, i.e., when a population of 10 animals was housed in twice as much space as a population of 5 animals and in 10 times the space of individually housed animals.

Novelty

A few studies have been concerned with the effect of stimulus change on endocrine response, on the assumption that a larger population offers the possibility of greater novelty and that novelty might be the important variable in the density studies. Christian and Davis (1955) tested the possibility that density reduction might be as stressful as density expansion. Rat populations in three Baltimore city blocks were reduced by trapping to about one-half of their estimated maximum and were maintained at that level for several months. An over-all reduction rather than increase in adrenal weight was found, suggesting that the population reduction was not stressful, at least as measured by changes in adrenal weight. It should be noted that possible transitory endocrine changes immediately following the density reductions were not measured.

A study by Siegel (1959c) also indicates that a density reduction is equivalent to a reduction in stress, as measured by adrenal weight regression. Twenty-five birds from each of two different groups of white Leghorn female chickens, housed 50 and 150 birds per pen, were sacrificed over a 15-day period. As expected, adrenal hypertrophy was more extensive in birds coming from the larger group. In both populations adrenal weights were significantly related to the day of sacrifice, with regression equations indicating that adrenal glands weighed progressively less as autopsies continued over the 15 day-period and population density decreased.

Vandenbergh (1960) found a transitory drop in eosinophil count following grouping of mice. This response, indicative of increased adrenocortical secretion, reached a peak approximately 4 hours after grouping and had largely disappeared by the second week. Change in adrenal weight was less rapid and less transitory. Christian (1959a) found an initial increase in urinary corticosteroid levels in guinea pigs following grouping, followed by a return to pregrouping levels within 3 days. Other investigators (Holcomb, 1957; Levine, 1959; Mason, 1959; Vogt, 1951) have found that almost any shift in stimulation will alter eosinophil and corticosteroid levels. It may therefore be that either a density increase or a density decrease would result in an initial rise in corticoid secretion, whereas only an increase in density would result in noticeable adrenal hypertrophy and other gross morphological changes. The possibly transitory stimulating effect of density reduction has not as yet been demonstrated, however. Some evidence that novelty is not the crucial factor in the more persistent morphological changes associated with high density is the differentially greater response of the low-status animals (see previous discussion). It seems probable that the dominant animals encounter as many or more novel situations as do the more socially restricted low-status animals, and yet their glandular response is less.

Effect of Tranquilizers

One study has been done on the effect of tranquilizers on endocrine response to population density (Christian, 1956b). Mice receiving reserpine in their drinking water showed less extensive glandular alteration than did similarly grouped mice not receiving tranquilizer. The results are interpreted as supporting the hypothesis that the density-related changes in endocrine function are due to sociophysiological response to group pressures.

Female Estrus Cycle

Several studies have focused on the effect of population density on the female estrus cycle. van der Lee and Boot (1955, 1956) found that housing female mice four to a cage often prolonged by several days the normal 4 to 6 day occurrence of estrus. This temporary suspension of estrus of grouped females was confirmed by Dewar (1959), Lamond (1958, 1959), and Whitten (1956, 1957, 1958, 1959). Whitten (1959) reports suspension of estrus for as long as 40 days when females are grouped 30 to a cage, with the estrus cycles promptly returning when the mice are separated into individual cages.

These results suggest the possibility that prolonged female diestrus occurs in dense populations and is one mechanism of density control. There is, however, evidence that tends to contradict such a conclusion. Whitten (1956, 1957, 1959) and Lamond (1959) have demonstrated that the introduction of a male into the female group or that the placing of a previously grouped female with a male will terminate the diestrus and will usually lead to pregnancy in a few days. Mating occurred predominantly on the third night after pairing when previously grouped females were placed with a male, indicating that contact with the male terminated a diestrus period and initiated an estrus cycle (Whitten, 1956, 1959). In contrast, matings with females previously housed individually were more randomly distributed among the first four nights, indicating the pre-existence of estrus cycles unrelated to the introduction of the male.

The ability of a male to terminate the diestrus of grouped females and the absence to date of reports of observed density-related increase in diestrus of females in mixed populations suggest that it is not a predominant factor in population control. The evidence nevertheless is consistent that the grouping of females results in diestrus, and this effect may be an important factor in determining the endocrinological or behavioral status of subjects used in laboratory settings. Whitten (1959) posits that some of the effects observed are mediated by the pituitary-gonadotrophic function, a theory that would relate these results closely to other observed effects of population density on endocrine function.

Controversy exists concerning the nature of the diestrus of the grouped females. Some investigators (Dewar, 1959; van der Lee & Boot, 1955, 1956) have attributed the diestrus to pseudopregnancy. Others (Lamond, 1959; Whitten, 1956, 1957, 1958) consider the condition to differ in crucial respects from true pseudopregnancy, being easily terminable at any time by the introduction of a male, being associated with reduced weight of ovaries and uterus and with reduced number or absence of corpora lutea, and being accompanied by mucified vagina. With the possible exception of mucified vagina, none of these effects would be expected with true pseudopregnancy (Nalbandov, 1958; Turner, 1955).

Whitten (1957) argues that severe stress reactions are not present in the grouped females, since they appear healthy, retain their body weight, return to estrus rapidly upon isolation or pairing with a male, and become pregnant without apparent difficulty. Christian (1960) cites recent evidence, however, suggesting some endocrine response to grouping of females. As compared to isolated controls, he found mild hyperplasia of the adrenal fasciculata-reticularis zone in grouped females, suggesting increased ACTH production by the pituitary. He also cites other evidence suggestive of increased pituitary-adrenal response. The response was not, however, so great as that observed in groupings of males or of mixed sexes.

Present evidence suggests that olfactory cues are important in initiating diestrus in grouped females, in terminating diestrus, in controlling sexual behavior leading to pregnancy, and even in preserving or disrupting pregnancy after it occurs. Lamond (1958) and Whitten (1959) found that females housed singly but separated from each other only by a partition showed disruption of the estrus cycle. van der Lee and Boot (1956) found that olfactory bulb removal reduced the number of females that became diestrus under conditions of grouping. Whitten (1956) found that mating of a grouped female could be shifted predominantly to the first night, instead of the third night, following pairing with a male if a male were enclosed within a small basket in the female cage for the 2 days prior to pairing or if the females were placed in a cage recently contaminated by males. Lamond (1959) reports that the number of litters born to anosmic mice is significantly smaller than for either normal or blinded animals. Bruce and Parrott (1960) report that pregnancy is blocked in a high proportion of recently mated intact female mice exposed to strange males, but not in anosmic females so exposed. Whether or not the olfactory cues that appear to mediate these effects operate through an effect on the pituitary-adrenal-gonadal system has not yet been established.

Effects on Nonreproductive Behavior

If density of population affects endocrine function, then it will almost inevitably affect behavior studies. The relationships between population density and behavior will not be reviewed, other than briefly to indicate that the effects may be crucial for many studies. With regard to learning ability, for example, Marx (1956) found that grouped rats could learn a vigorous lever-pressing response faster than individually housed animals. With regard to "emotionality," Bovard and Newton (1956) found that rats living in a group showed more defecation and vocalization when transported by the experimenter. Much work has appeared and is appearing on the effects of early handling on later behavior. The evidence (e.g., Levine, 1959) that the early handling effects are mediated by endocrine response to a "stressful" situation suggests the important role that endocrine function, and therefore population density, may have on such diverse variables as learning ability, survival, and brain chemistry. Clearly, the field is ripe for more experimental work. Also, clearly, the experimenter who draws his subjects haphazardly from colony cages containing varying numbers of animals is introducing into his study an uncontrolled variable that may crucially affect his obtained results.

Summary

Population density has been shown to affect endocrine function, being positively related to adrenal hypertrophy and adrenocortical activity and negatively related to gonadal and mammary activity. Other factors being equal, many reactions appear to vary as the logarithm of the population size. However, there is some evidence that marked genetic differences in response exist as well as some evidence that population size is only indirectly a causative agent. Amount of wounding does not appear to be crucially important, although social rank, which at times correlates with amount of wounding, is a good predictor of individual response to population pressures. Mechanical restriction of living space appears to be unimportant within broad limits. Response to novelty may account for some transitory endocrine reactions but seems unlikely to be the crucial variable in less transitory morphological effects of population size. Grouping effects on female estrus cycle appear related more to olfactory cues and sexual composition of the group than to population density per se. One effect of the endocrine response appears to be an alteration

of reproductive capacity such as to provide a self-limiting control of population size. Learning ability, emotionality, and other behavior may also be altered by variations in density of population.

REFERENCES

BARNETT, S. A. Competition among wild rats. *Nature, Lond.*, 1955, **175**, 126.

BEACH, F. A. *Hormones and behavior.* New York: Hoeber, 1948.

BINDRA, D. *Motivation: A systematic reinterpretation.* New York: Ronald, 1959.

BOVARD, E. W., & NEWTON, D. G. Systematic early handling and prolonged experience with the mother as developmental variables in the male albino rat. In N. R. Brewer (Ed.), *Proceedings of animal care panel.* Chicago: Animal Care Panel, 1956. Pp. 67–74.

BRUCE, H. M., & PARROTT, D. M. V. Role of olfactory sense in pregnancy block by strange males. *Science*, 1960, **131**, 1526.

CALHOUN, J. B. The social aspects of population dynamics. *J. Mammal.*, 1952, **33**, 139–159.

CHITTY, D. Adverse effects of population density upon the viability of later generations. In J. B. Cragg & N. W. Pirie (Eds.), *The numbers of man and animals.* Edinburgh: Oliver & Boyd, 1955. Pp. 57–67.

CHITTY, D., CHITTY, HELEN, LESLIE, P. H., & SCOTT, J. C. Changes in the relative size of the nucleus in the intervertebral discs of stressed Orkney voles (*Microtus oreadensis.*) *J. Pathol. Bacteriol.*, 1956, **72**, 459–470.

CHRISTIAN, J. J. The adreno-pituitary system and population cycles in mammals. *J. Mammal.*, 1950, **31**, 247–259.

CHRISTIAN, J. J. Effect of population size on the adrenal glands and reproductive organs of male white mice. *Amer. J. Physiol.*, 1955, **181**, 477–480. (a)

CHRISTIAN, J. J. Effect of population size on the weights of the reproductive organs of male white mice. *Amer. J. Physiol.*, 1955, **182**, 292–300. (b)

CHRISTIAN, J. J. Adrenal and reproductive responses to population size in mice from freely growing populations. *Ecology*, 1956, **37**, 258–273. (a)

CHRISTIAN, J. J. Reserpine suppression of density-dependent adrenal hypertrophy and reproductive hypoendocrinism in populations of male mice. *Amer. J. Physiol.*, 1956, **187**, 353–356. (b)

CHRISTIAN, J. J. A discussion of a paper by J. W. Mason, Psychological influences on the pituitary-adrenal cortical system. *Recent Progr. hormone Res.*, 1959, **15**, 345–389. (a)

CHRISTIAN, J. J. Lack of correlation between adrenal weight and injury in grouped male albino mice. *Proc. Soc. Exp. Biol. Med.*, 1959, **101**, 166–168. (b)

CHRISTIAN, J. J. The role of endocrine and behavioral factors in the growth of mammalian populations. In Gorbman (Ed.), *Comparative endocrinology.* New York: Wiley, 1959. Pp. 71–97. (c)

CHRISTIAN, J. J. Adrenocortical and gonadal responses of female mice to increased population density. *Proc. Soc. Exp. Biol. Med.*, 1960, **104**, 330–332.

CHRISTIAN, J. J., & DAVIS, D. E. The reduction of adrenal weight in rodents by reducing population size. *Trans. 20th N. Amer. Wildlife Conf.*, 1955, 177–189.

CHRISTIAN, J. J., & DAVIS, D. E. The relationship between adrenal weight and population status of urban Norway rats. *J. Mammal.*, 1956, **37**, 475–486.

CHRISTIAN, J. J., & LEMUNYAN, C. D. Adverse effects of crowding on reproduction and lactation of mice and two generations of their progeny. *Endocrinology*, 1958, **63**, 517–529.

CLARKE, J. R. The effect of fighting on the adrenals, thymus and spleen of the vole (*Microtus agrestia*). *J. Endocrinol.*, 1953, **9**, 114.

CREW, F. A., & MIRSKAIA, L. Effect of density on adult mouse populations. *Biol. gen.*, 1931, **7**, 239–250.

DAVIS, D. E., & CHRISTIAN, J. J. Relation of adrenal weight to social rank in mice. *Proc. Soc. Exp. Biol. Med.*, 1957, **94**, 728–731.

DEWAR, A. D. Observations on pseudopregnancy in the mouse. *J. Endocrinol.*, 1959, **18**, 186–190.

HELMREICH, R. L. Regulation of reproductive rate by intra-uterine mortality in the deer mouse. *Science*, 1960, **132**, 417–418.

HOFFMAN, R. S. The role of reproduction and mortality in population fluctuations of voles (*Microtus*). *Ecol. Monogr.*, 1958, **28**, 79–109.

HOLCOMB, R. B. Investigations on the urinary excretion of "reducing corticoids" in cattle and sheep. *Acta endocrinol. Suppl., Kbh.*, 1957, **34**, 1–100.

KALELA, O. Regulation of reproduction rate in subarctic populations of the vole *Clethrionomys rufocanus* (Sund.). *Ann. Acad. Sci. Fennicae, Ser. A*, 1957, **34**, 1–60.

LAMOND, D. R. Spontaneous anoestrus in mice. *Proc. Aust. Soc. Anim. Prod.*, 1958, **2**, 97–101.

LAMOND, D. R. Effect of stimulation derived from other animals of the same species on oestrous cycles in mice. *J. Endocrinol.*, 1959, **18**, 343–349.

LEVINE, S. The psychophysiological effects of infantile stimulation. Paper read at American Association for the Advancement of Science, Chicago, December 1959.

LOUCH, C. D. Adrenocortical activity in relation to the density and dynamics of three confined populations of *Microtus pennsylvanicus*. *Ecology*, 1956, **37**, 701–713.

MARX, M. H. Some relations between frustration and drive. In M. R. Jones (Ed.), *Nebraska symposium on motivation: 1956.* Lincoln: Univer. Nebraska Press, 1956. Pp. 92–130.

MASON, J. W. Psychological influences on the pituitary-adrenal cortical system. *Recent Progr. hormone Res.*, 1959, **15**, 345–389.

NALBANDOV, A. V. *Reproductive physiology.* San Francisco: Freeman, 1958.

SELYE, H. The general adaptation syndrome and the diseases of adaptation. *J. clin. Endocrinol. Metab.*, 1946, **6**, 117–230.

SIEGEL, H. S. Egg production characteristics and adrenal function in white Leghorns confined at different floor space levels. *Poult. Sci.*, 1959, **38**, 893–898. (a)

SIEGEL, H. S. The relation between crowding and weight of adrenal glands in chickens. *Ecology*, 1959, **40**, 494–498. (b)

SIEGEL, H. S. Effect of population density on the pituitary-adrenal cortical axis of cockerels. *Poult. Sci.*, 1960, **39**, 500–510.

SOUTHWICK, C. H. The population dynamics of confined house mice supplied with unlimited food. *Ecology*, 1955, **36**, 212–225. (a)

SOUTHWICK, C. H. Regulatory mechanisms in house mouse populations: Social behavior affecting litter survival. *Ecology*, 1955, **36**, 627–634. (b)

SOUTHWICK, C. H., & BLAND, V. P. Effect of population density on adrenal glands and reproductive organs of CFW mice. *Amer. J. Physiol.*, 1959, **197**, 111–114.

STERKER, R. L., & EMLEN, J. T. Regulatory mechanisms in house mouse populations: The effect of limited food supply on a confined population. *Ecology*, 1953, **34**, 375–385.

TURNER, C. D. *General endocrinology.* Philadelphia: Saunders, 1955.

VANDENBERGH, J. G. Eosinophil response to aggressive behaviour in CFW albino mice. *Anim. Behav.*, 1960, **8**, 13–18.

VAN DER LEE, S., & BOOT, L. M. Spontaneous pseudopregnancy in mice. *Acta physiol. pharmacol. Neerl.*, 1955, **4**, 442–444.

VAN DER LEE, S., & BOOT, L. M. Spontaneous pseudopregnancy in mice. II. *Acta physiol. pharmacol. Neerl.*, 1956, **5**, 213–215.

VOGT, M. The effect of emotion and of beta-tetrahydronaphthylamine on the adrenal cortex of the rat. *J. Physiol.*, 1951, **114**, 465–470.

WHITTEN, W. K. Modification of the oestrous cycle of the mouse by external stimuli associated with the male. *J. Endocrinol.*, 1956, **13**, 399–404.

WHITTEN, W. K. Effect of exteroceptive factors on the oestrous cycle of mice. *Nature, Lond.*, 1957, **180**, 1436.

WHITTEN, W. K. Modification of the oestrous cycle of the mouse by external stimuli associated with the male. *J. Endocrinol.*, 1958, **17**, 307–313.

WHITTEN, W. K. Occurrence of anoestrus in mice caged in groups. *J. Endocrinol.*, 1959, **18**, 102–107.

(Received August 10, 1960)

Psychological Bulletin
1961, Vol. 58, No. 6, 452-458

DOES THE HEART LEARN?

DONALD SHEARN
Colorado College

In our time of cohabitation of various sciences one may wonder about the kind of affairs psychology has with some of the more firmly established disciplines. Other sciences may very well believe that the progeny of a relationship with psychology would necessarily be illegitimate. Or perhaps, at best, that psychology would have everything to gain and nothing to give. Must psychology be the protegee, or does it have unique techniques to share? It is the aim of this paper, by way of presenting some experimental findings, to suggest that certain techniques of modern psychology can be useful in the analysis of problems of cardiovascular physiology.

Detailed encouragement to the psychologist to use his techniques in the physiological laboratory comes from current adjustments in physiological thinking. Reviews of recent circulatory research by Rushmer (1955), Rushmer and Smith (1959), point out that the cardiovascular system is not altogether faithful to a few classical laws based largely upon simple hydraulic principles (Bainbridge, 1915; Patterson, Piper, & Starling, 1914). Rather it appears that this system is true to many principles roving about on several levels of analysis, among them the principles derived from conditioning procedures.

Gantt's translation of *The Internal Organs and the Cerebral Cortex* by Bykov (1957) could be the signal for a methodological revolution in certain phases of biological analysis and control, wherein the physiological reactions of the intact organism are modified by conditioning techniques. Of particular interest in this book for our present discussion is the chapter on circulatory adjustments. Several experiments are cited wherein cardiac, vasomotor, and even splenic responses are conditional upon exteroceptive stimuli presented by the experimenter.

In contrasting the results mentioned in the Bykov book with those from other sources we are brought to what is perhaps the most paradoxical feature of cardiovascular conditioning, the form of the conditioned cardiac response. Granting that the heart does learn, just what is it that is learned?

What Does the Heart Learn?

CR-UCR Similarity. Soviet investigators generally suggest that the conditioned response (CR) closely resembles the unconditioned response (UCR) as illustrated by an experiment of Petrova (Bykov, 1957). An auditory stimulus (whistle) was combined with intravenous injections of nitroglycerin. Because the act of injecting the fluid would act as a conditioned stimulus, its effect was extinguished with repeated intravenous injections of normal saline. The whistle, on the other hand, was always sounded after the nitroglycerin had been injected (but before the effect of the drug was manifest). After about 100 pairings of the whistle and nitroglycerin the whistle presented alone produced changes typical of those elicited by the drug (accelerated heart rate, decrease in QRS voltage, and augmented P and T waves).

Delov (Bykov, 1957) demonstrated that a conditioned stimulus may produce a very different response from the above when combined a number of times with a drug of different consequences. In this experiment the conditioned stimulus (CS) was actually the stimulus complex associated with the injection, while the unconditioned stimulus (UCS) was a 0.2-gram injection of morphine. After 20 to 30 injections, the CS given without morphine produced the same changes in the electrocardiogram as those produced by morphine (deceleration in heart rate and marked reduction in the P deflection).

Additional experiments (Bykov, 1957) showing the similarity of UCR and CR for other drugs have been conducted by Samarin (strophanthin) and Levitin (acetylcholin and epinephrine).

Other investigators have taken a different view of the form of the heart rate CR. For example, in some experiments with human subjects by Zeaman, Deane, and Wegner (1954) and Zeaman and Wegner (1954), it was suggested that the CR resembles the UCR at the time of the UCS (shock) termination. In accord with this hypothesis a 2-second shock gave an accelerated heart rate CR and a 6-second shock gave a decelerated CR (since the UCR at shock termination was accelerating or decelerating, respectively). When other shock values were used in a later experiment (Zeaman & Wegner, 1958) this hypothesis was not upheld. It was predicted from the hypothesis that no conditioning would occur for a very short shock (0.1 second) which did not allow a change in heart rate before its termination, or for a very long shock (15 seconds) which allowed the heart rate to return to normal by the time it was terminated. When conditioning did occur these investigators revised their hypothesis to suggest that to some extent large UCRs tend to give accelerative CRs and small UCRs decelerative CRs.

Decelerative CR. A decelerative heart rate CR in human subjects is consistently reported by Bersh, Notterman, and Schoenfeld (1953, 1956a, 1956b, 1956c, 1957a, 1957b; Notterman, Schoenfeld, & Bersh, 1952a, 1952b, 1952c). Their procedure was essentially the same as that used by Zeaman and Wegner (1954) with which a decelerative CR was obtained (1-second CS, 6-second CS-UCS interval, and a 6-second UCS). Their UCS shock level, however, was over twice that of the Zeaman and Wegner studies (30-volt alternating current as contrasted with 13-volt alternating current). For the most part, the measures indicating a decelerative CR were taken during the last two heart cycles of the CS-UCS interval. In answer to possible criticism that deceleration during this portion of the interval was not a representative CR, they also measured the first two heart cycles of the CS-UCS interval (Notterman et al., 1952c) and again found a decreasing heart rate, which was not, however, statistically significant.

Owens and Gantt (1950) report a decelerative CR when the petting of a dog served as the UCS. The UCR to this stimulation was also a reduction in heart rate. Mixed results regarding the form of the heart rate CR were obtained by Beier (1940). One subject showed an accelerative CR, another a decelerative CR, and still another, conditioned arrhythmia. The UCS used in this experiment was the working of a bicycle ergometer by the subject.

Accelerative CR. Other experiments

indicate a CR which is predominately accelerative in form. Skaggs (1926) used an auto horn CS and an induction shock UCS, separated by 1 minute, to produce a mild increase in human heart rate (1.1 beats/minute). A greater increase in rate was observed between the "normal" condition and the "expectancy" period preceding the CS (9.4 beats/minute).

Anderson and Parmenter (1941) demonstrated that the CR is an increase in heart rate when a buzzer or metronome CS is used with a shock pulse UCS. They further demonstrated "neurosis" in their sheep subjects with a discrimination procedure where only one of two stimuli was paired with shock. Neurotic subjects showed a higher and more irregular heart rate than normals in the experimental room, and gave an increase in heart rate to incidental stimuli whereas normals did not.

Moore and Marcuse (1945) ran two sows daily for 10 months using a tone CS and food UCS. They found a reliable increase in heart rate upon presentation of the CS, which preceded the UCS by 1 minute. Dykman and Gantt (1951) used a tone CS and a shock UCS, separated by 2 minutes, to produce an accelerative CR in dogs. As noted earlier, Zeaman and Wegner (1954), using a 2-second shock UCS, showed an increase in heart rate in human subjects with onset of the CS.

CS-UCS Interval and Regularity. Church and Black (1958) using dog subjects also found an accelerative CR with a tone CS and a 3-second shock UCS. Their results indicate that CR latency is shorter for a 5-second CS-UCS interval than for a 20-second CS-UCS interval. Latencies were virtually the same for the trace and delay conditioning procedures. No substantial differences in heart rate were observed between the various experimental treatments. This last finding is to be contrasted with some results of Bersh, Notterman, and Schoenfeld (1953) who found that an irregular time between CS and UCS produced more "anxiety" (i.e., heart rate CRs of greater magnitude) than a regular time between. A condition where shock did not always follow the CS produced more anxiety than either of these conditions.

Resistance to Extinction. One particular disclosure from the Soviet cardiac conditioning work seems to be of special importance (Bykov, 1957). That is, the CR developed in pairing a neutral stimulus with a pharmacological agent is very hard to extinguish. For example, some 296 presentations of the CS alone were required by Petrova to extinguish the cardiac CR.

Gantt (Bykov, 1957) reports that a cardiac conditioned reflex to food may persist 2 years after the salivary and motor components have been extinguished. Notterman, Schoenfeld, and Bersh (1952c) found that irregular pairing of the UCS with the CS gave greater resistance to extinction than regular reinforcement. They report further (1952a) that when subjects could avoid the shock UCS with a skeletal response, extinction was more rapid than when subjects were told there would be no shock in extinction. Both of these treatments produced more rapid extinction than the regular extinction procedure. In a later experiment (Bersh et al., 1956c) found that the CRs of subjects who were forcibly restrained from making the skeletal avoidance response extinguished more rapidly than the free avoidance subjects.

Generalization. Stimulus generalization of the CS has been demonstrated by Dykman and Gantt (1951) whose dog subjects differentiated between 256, 512, and 1,024 cps tones with respect to heart rate, latency, and EKG amplitude. Bersh, Notterman, and Schoenfeld (1956c) obtained generalization across tone frequencies as a function of intensity of the UCS. For a 28-volt alternating current shock UCS the human *S*s showed a greater CR (depression of rate) and a flatter generalization across the 1,920, 1,020, 480, and 180 cps tones than for a 20-volt alternating current UCS.

CR across Trials. Heart rate conditioning data collected by Dawson (1953) are perhaps the best source of information for changes in the form of the heart rate CR across trials. They also illustrate sharply how deceptive a simple label such as a rate "increase" or "decrease" is in describing the cardiac CR. So far as such details are reported, most of the studies discussed earlier involved no more than 11 conditioning trials (e.g., Church & Black, 1958; Notterman et al., 1952b, 1952c; Zeaman & Wegner, 1954, 1958). In the Dawson experiment 20 conditioning trials were used and the second by second forms of the CR and UCR are shown for each five-trial block. These results show that the early CR is, in effect, an acceleration followed by a deceleration to the level preceding the CS. At this stage of conditioning a comparison of rates preceding and following the CS will show a net increase no matter which point within the CS-UCS interval is selected. As conditioning trials continue, however, the decelerative phase of the CR becomes more pronounced, such that rate of the heart cycles during this phase is less than the rate preceding the CS. Hence, increase or decrease in heart rate as *the* CR depends heavily upon the location within the CS-UCS interval one uses, as well as the trial number (or number of trials if trials are averaged). We may then add these factors to others which affect the CR, such as UCS length, CS-UCS interval, and the kind and intensity of the UCS.

Interaction with Other Body Systems

A factor which appears in elementary physiology texts suggests that the heart, as such, may not learn at all. This factor, obvious enough perhaps to be invisible, is respiration. Recent quantitative data clearly show how breathing may affect heart rate (Clynes, 1960; Huttenlocher & Westcott, 1957). Both inspiration and expiration produce a biphasic cardiac response: a brief accelerative phase followed by a decelerative phase of longer duration. This biphasic cardiac response is of greater magnitude and has a shorter latency for inspiration than for expiration (Clynes, 1960). Furthermore, it has been demonstrated that in a classical conditioning situation involving buzzer and shock, conditioned deep inspirations occur with the onset of the CS, and that the cardiac CR is a brief acceleration followed by a more pronounced deceleration (Huttenlocher & Westcott, 1957).

Regardless of which portion of the respiratory cycle might be correlated with the CS, there is the frightful prospect that cardiac conditioning work thus far has, in fact, been unknowingly concerned with respiratory conditioning. Or, with some luck, cardiac conditioning has merely been contaminated by the respiratory variable.

Fortunately, at least one cardiac conditioning experiment has been reported in which respiration was controlled. Westcott (1959) instructed subjects to breathe shallowly in time with a metronome during 10 CS (buzzer) alone trials and 10 conditioning trials when the CS and UCS (shock) were paired. The cardiac response in this experiment was a net drop in rate when the CS was given before conditioning, and a net increase in rate after the second conditioning trial. The conditioning curve was negatively accelerated across trials, showing an increase in heart rate over the pre-CS rate of 3.2 beats per minute on the last two trials. Respiration records showed consistent breathing on each trial, and across trials, for both frequency of respiration and the I/E ratios.

There are still other doubts about cardiac conditioning which we should consider. Kendon Smith (1954) argues that all conditioned visceral responses are in reality artifacts because they are brought on by activation of the skeletal musculature. According to this reasoning innate neural connections from the skeletal muscles activate the visceral systems with a muscular "bracing" to the UCS. Skeletal reactions are said to provide numerous afferent cues whereas "autonomic reactions generate no regulatory feedback whatever." Hence it is the skeletal system which is conditioned and the visceral system which merely accompanies. These ideas do badly in finding support from the cardiac literature discussing afferent pathways (e.g., Mitchell, 1956; Rushmer & Smith, 1959) for there is considerable anatomical evidence for autonomic feedback from the carotid and aortic bodies.

The opposite hypothesis, that skeletal responses can be mediated by autonomic responses is suggested by Wenzel (1959). Her data show an increase in heart rate to tones associated with food and a decrease to tones associated with shock. Whether or not heart rate differentiation between the two conditions is related to autonomic mediation of skeletal responses is yet to be shown in the laboratory.

Church and Black (1958) argue a similar case which is in line with Pavlov's "inhibition of delay." They too suggest autonomic mediation of skeletal responses. The tabulated latencies in their experimental report, however, tend to show shorter skeletal than autonomic latencies, which, after all, are consistent with the time constants of the two systems.

Perhaps the complexity of the organism is such as to preclude such simple cause and effect hypotheses about the various systems.

Conclusion

That the activity of the heart will change significantly in amplitude and rate in the presence of conditional stimuli is clear enough. There appear, however, to be mixed emotions as to the form of the heart rate CR since some authors report an increased rate, others a decreased rate, and still others either increased or decreased rate, depending on such factors as the UCR. The original question might then be what does the heart learn rather than does it learn? It is suggested that an answer to the second question may be found at two locations: at the desk, where heart rate changes would be treated as analog events rather than simple up or down events, and, at the laboratory, where like problems have previously been untangled with parametric study.

Some tentative principles have been abstracted from the papers reviewed:

1. Both form of the EKG cycle

and heart rate may be conditioned with the classical paradigm.

2. Latency of the heart rate CR is less for a shorter CS-UCS interval.

3. A CR of greater magnitude is produced when the UCS irregularly follows the CS.

4. CR resistance to extinction is great when some pharmacological UCSs are used. Resistance to extinction is increased with irregular CS-UCS pairings, and is decreased when UCS avoidance is made contingent upon a skeletal response.

5. There is a generalization gradient across tone frequencies as a function of UCS intensity.

6. The heart rate CR changes across trials such that the dominant accelerative portion of the response decreases as the decelerative portion increases.

Whether it is "really" the heart that learns, or something else such as the respiratory or the skeletal system, is perhaps a matter of degree. It seems unlikely that a particular bodily system is completely free from the influence of other bodily systems.

REFERENCES

Anderson, O. D., & Parmenter, R. A long term study of the experimental neurosis in the sheep and dog. *Psychosom. med. Monogr.*, 1941, **2**(3–4).

Bainbridge, F. A. The influence of venous filling upon the rate of the heart. *J. Physiol.*, 1915, **50**, 65–84.

Beier, C. D. Conditioned cardiovascular responses and suggestions for treatment of cardiac responses. *J. exp. Psychol.*, 1940, **26**, 311–321.

Bersh, P. J., Notterman, J. M., & Schoenfeld, W. N. The effect of randomly varying the interval between conditioned and unconditioned stimuli upon the production of experimental anxiety. *Proc. Nat. Acad. Sci., Wash.*, 1953, **39**, 553–570.

Bersh, P. J., Notterman, J. M., & Schoenfeld, W. N. Extinction of human cardiac response during avoidance conditioning. *Amer. J. Psychol.*, 1956, **69**, 244–251. (a)

Bersh, P. J., Notterman, J. M., & Schoenfeld, W. N. Generalization to varying frequencies as a function of intensity of unconditioned stimulus. *USAF Sch. Aviat. Med. Rep.*, 1956, No. 56–79. (b)

Bersh, P. J., Notterman, J. M., & Schoenfeld, W. N. Relations between acquired autonomic and motor behavior during avoidance conditioning. *USAF Sch. Aviat. Med. Rep.*, 1956, No. 56–80. (c)

Bersh, P. J., Notterman, J. M., & Schoenfeld, W. N. A comparison of internal vs. external reinforcement in motor avoidance situations. *USAF Sch. Aviat. Med. Rep.*, 1957, No. 57–27. (a)

Bersh, P. J., Notterman, J. M., & Schoenfeld, W. N. The efficiency of pursuitrotor performance during experimentally induced anxiety. *USAF Sch. Aviat. Med. Rep.*, 1957, No. 57–28. (b)

Bykov, K. M. *The cerebral cortex and the internal organs.* (Ed. & Trans. by W. H. Gantt) New York: Chemical Publishing, 1957.

Church, R. M., & Black, A. H. Latency of the conditioned heart rate as a function of the CS-US interval. *J. comp. physiol. Psychol.*, 1958, **51**, 478–487.

Clynes, M. Computer analysis of reflex control and organization: Respiratory sinus arrhythmia. *Science*, 1960, **131**, 300–302.

Dawson, H. E. Concurrent condition of autonomic processes in humans. Unpublished doctoral dissertation, Indiana University, 1953.

Dykman, R. A., and Gantt, W. H. A comparative study of cardiac responses and motor conditioned responses in controlled "stress" situation. *Amer. Psychologist*, 1951, **6**, 263. (Abstract)

Huttenlocher, J., & Westcott, M. R. Some empirical relationships between respiratory activity and heart-rate. *Amer. Psychologist*, 1957, **12**, 414. (Abstract)

Mitchell, G. A. G. *Cardiovascular innervation.* Edinburgh: Livingstone, 1956.

Moore, A. U., & Marcuse, F. L. Salivary, cardiac, and motor indices of conditioning in two sows. *J. comp. Psychol.*, 1945, **38**, 1–16.

Notterman, J. M., Schoenfeld, W. N., & Bersh, P. J. A comparison of three extinction procedures following heart rate conditioning. *J. abnorm. soc. Psychol.*, 1952, **47**, 674–677. (a)

Notterman, J. M., Schoenfeld, W. N., & Bersh, P. J. Conditioned heart rate response in human beings during experimental anxiety. *J. comp. physiol. Psychol.*, 1952, **45**, 1–8. (b)

NOTTERMAN, J. M., SCHOENFELD, W. N., & BERSH, P. J. Partial reinforcement and conditioned heart rate response in human subjects. *Science*, 1952, **115**, 77–79. (c)

OWENS, O., & GANTT, W. H. Does the presence of a person act on cardiac rate of the dog as unconditional stimulus? *Amer. J. Physiol.*, 1950, **163**, 746. (Abstract)

PATTERSON, S. W., PIPER, H., & STARLING, E. H. The regulation of the heart beat. *J. Physiol.*, 1914, **48**, 465.

RUSHMER, R. F. Applicability of Starling's law of the heart to intact, unanesthetized animals. *Physiol. Rev.* 1955, **35**, 138–142.

RUSHMER, R. F., & SMITH, O. A. Cardiac control. *Physiol. Rev.*, 1959, **39**, 41–68.

SKAGGS, E. B. Changes in pulse, breathing, and steadiness under conditions of startledness and excited expectancy. *J. comp. Psychol.*, 1926, **6**, 303–318.

SMITH, K. Conditioning as an artifact. *Psychol. Rev.*, 1954, **61**, 217–225.

WENZEL, B. M. Changes in heart rate associated with positive and negative reinforcement and their modification by reserpine. Paper read at the American Psychological Association, New York, September 1959.

WESTCOTT, M. R. The acquisition of a conditioned cardiac acceleration in humans. Paper read at Eastern Psychological Association, New York, 1959.

ZEAMAN, D., DEANE, G., & WEGNER, N. Amplitude and latency characteristics of the conditioned heart response. *J. Psychol.*, 1954, **38**, 235–250.

ZEAMAN, D., & WEGNER, N. The role of drive reduction in the classical conditioning of an autonomically mediated response. *J. exp. Psychol.*, 1954, **48**, 349–354.

ZEAMAN, D., & WEGNER, N. Strength of cardiac conditioned responses with varying stimulus durations. *Psychol. Rev.*, 1958, **65**, 238–241.

(Received July 15, 1960)

Psychological Bulletin
1961, Vol. 58, No. 6, 459–490

MATERNAL DEPRIVATION: TOWARD AN EMPIRICAL AND CONCEPTUAL RE-EVALUATION[1]

LEON J. YARROW
Family and Child Services, Washington, D. C.

The significance of early infantile experience for later development has been reiterated so frequently and so persistently that the general validity of this assertion is now almost unchallenged. An extensive literature on deviating patterns of maternal care, loosely labeled "maternal deprivation," adds up with an impressive consistency in its *general* conclusions: deviating conditions of maternal care in early life tend to be associated with later disturbances in intellectual and personal-social functioning. It has been difficult to build on this general premise in formulating more precise research hypotheses relating specific variables of early maternal care to later developmental characteristics. If one attempts to order the empirical data from the many studies and the varied contexts, it becomes apparent that the concept of maternal deprivation is a rather muddied one. Maternal deprivation has been used as a broad descriptive term as well as an overall explanatory concept. As a descriptive term it encompasses a variety of conditions of infant care which are phenotypically as well as dynamically very different. In this review of the research and theoretical literature, our major objective is to clarify the concept of maternal deprivation by identifying the basic variables and concepts which have been indiscriminately combined under this term.

Previous reviews have dealt primarily with the findings (Bowlby, 1951; Glaser & Eisenberg, 1956), or with the methodology of a few studies (Pinneau, 1950, 1955). The chief effort of this review will be directed towards sorting out on an empirical level the varied antecedent conditions of maternal care described in the literature, and relating these empirical conditions to some major theoretical concepts. Through this kind of analysis, it is hoped to facilitate the formulation of more explicit hypotheses on the relationship between specific aspects of early life experiences and later development.

EMPIRICAL ANALYSIS OF THE RESEARCH ON "MATERNAL DEPRIVATION"

In the literature on maternal deprivation, four different kinds of deviations from a hypothetical mode of maternal care have been included: institutionalization; separation from a mother or mother-substitute; multiple mothering, in which there is no one continuous person performing the major mothering functions; distortions in the quality of mothering, e.g., rejection, overprotection, am-

[1] Based on invited address, Division on Developmental Psychology, sixty-seventh Annual Convention of the American Psychological Association, Cincinnati, Ohio, September 1959.

This paper was prepared in conjunction with a research project on The Effects of a Change in Mother-Figure during Infancy on Personality Development, conducted under Research Grant 3M-9077 from the National Institute of Mental Health, United States Public Health Service.

bivalence. In very few studies do we find these "pure conditions." Most often several conditions occur concomitantly or sequentially in complex interaction, e.g., separation is followed by institutionalization, multiple mothering occurs in an institutional setting.

Tables 1 to 4 present the chief research studies organized in terms of the major conditions of early care: institutionalization, separation, multiple mothering. Studies on distortions in the mother-child relationship, e.g., rejection, overprotection, ambivalence, on which there are many clinical reports, but few research reports, have not been included. The studies presented in the tables are grouped according to their general research designs: retrospective, direct, or contemporaneous. The tables point out the major characteristics of the samples: the population from which the subjects were chosen, the ages at the time of study, and the ages at the time of the experience. Also presented are the major techniques used in data collection or the kinds of data obtained. For the retrospective studies, the presence or absence of data on earlier conditions of maternal care is noted. Finally, overlapping or contaminating conditions are noted where they have been reported.

It is clear from the tables that the major share of studies has been on institutional care. There are many fewer published reports on separation and multiple mothering. In the following section, in considering each of these types of studies, our focus will be on an analysis of the environmental conditions and the impact of these events and conditions on development. Throughout we will attempt to integrate the empirical data in terms of some basic psychological concepts, and to point up some hypotheses amenable to research.

INSTITUTIONALIZATION

Most of the generalizations about the effects of "maternal deprivation" are based on retrospective research in which institutionalization has been a major background condition. The general research designs of the many retrospective studies reported between 1937 and 1955 are basically similar and tend to suffer from similar methodological deficiencies. In all but a few studies there is a sampling bias due to the method of selection of cases; subjects are chosen from clinic populations of cases under treatment for emotional or personality disturbances. (In delving back into the history of these patients, it was discovered that many had spent some part of their earlier life in an institutional setting.) Perhaps the most significant deficiency in many of these studies is the lack of specific data on early conditions of maternal care. The characteristics of the institutional environment are unknown or not described, and no data, or, at the best, very meager data are given about the circumstances associated with institutionalization. Such significant information as age at time of placement, duration of institutional care, traumatic conditions preceding or concomitant with institutional placement is rarely given. Frequently information about experiences following institutional care is scant and of uncertain validity. The data on the personality characteristics of the subjects also vary greatly in depth and adequacy; much data are derived from psychiatric diagnoses based on an unspecified number of interviews or consist of case history material from unspecified sources; in a few instances, projective or other kinds of personality tests have been used.

The Institutional Environment

In much of the research on institutions the environment has been dealt with so grossly that "institutionalization" has often referred to a setting as broad in many respects as "the home." Only a few contemporaneous studies of infants and young children give sufficiently detailed descriptions of the institutional setting to enable one to isolate discrete variables. Only one study, comparing the institutional and home environments of a small group of infants, makes a serious attempt to give an objective description of an institution (Rheingold, 1960).

The institutional environments in the direct studies can be ordered in terms of several theoretically meaningful categories which can be further reduced to specific research variables.

The physical environment—quality and amount of sensory stimulation. The importance of sensory stimulation for development has recently been emphasized by a number of animal experiments. In most of the research, institutional settings are characterized in the extreme as lacking in sensory stimulation; they are described as colorless and drab with little visual or auditory stimulation and with few objects for the child to manipulate.

The emotional environment—affective stimulation. For research, the emotional environment can be defined in a restricted sense in terms of formal, measurable aspects of affective stimulation, i.e., intensity and variability. Institutions tend to be characterized by an emotional blandness and a lack of variation in feeling tone with the result that the infant is not exposed to strongly negative or strongly positive affective stimulation.

The social environment—social stimulation. The amount of mothering, the quality and consistency of mothering, and the amount and quality of general social stimulation are major aspects of the animate environment in terms of which institutional care is defined. Most of the studies describe a low adult-child ratio, averaging about one adult to 10 infants in institutional settings. There are usually many different caretakers, with the result that the infant has little opportunity to relate to one person as a consistent source of gratification. Compared with an infant in his own home, the research indicates that in institutions there is much less mothering contact, less total social stimulation, and less stability in mother-figures.

Learning conditions. Learning conditions which deviate from those in a "normal" home environment are reported characteristic of institutions: deviations in opportunities for acquiring or practicing new skills, deviations in motivational conditions, and in scheduling. Often infants are confined to the crib or playpen during most of the day, with very limited opportunity to practice emerging motor skills or to make perceptual discriminations. There tends to be little recognition by adults for positive achievements, with no or inconsistent reinforcement for positive learnings or socially desirable responses. Daily routines are sometimes characterized by an element of unpredictability, but more often routines are rigidly scheduled with little variation from day to day, and with little adaptation to individual differences.

It is clear that institutionalization

TABLE 1

Research on Institutionalization: Direct Studies of Children in Institutions

Investigator	Subjects	Age at time of study	Age when institutionalized	Techniques or type of data	Description of environment
Brodbeck & Irwin (1946)	Institutional: 94 Controls: 217	Birth to 6 months	Birth to 6 months	Analysis of speech sounds	General—social, emotional
Brown (1937)	Institutional: 200 Controls: 200 from "poor" home environments	9–14 years	Broad range: birth to adolescence	Brown Personality Inventory	No data
Dennis & Najarian (1957)	Institutional: 49 infants; 30 preschool age Controls: 41	2–12 months 4½–6 years	Birth	Infant tests Goodenough Draw-A-Man Test	Detailed—physical, social, learning conditions
DuPan & Roth (1955)	Institutional: 14	4–30 months	Birth to 3 months	Gesell test	Detailed—physical, social, learning conditions
Fischer (1952)	Institutional: 62	6–7 months	Birth to 3 months	Cattell test Observation	Detailed—physical, social
Flint (1957)	Institutional: 16	2–20 months	Birth to 6 months	Infant security scale Observation	Detailed—physical, social, learning conditions
Freud & Burlingham (1944)	Institutional: approximately 90	Birth to 2 years Longitudinal	Early infancy—no specific data	Clinical observation	General—physical, social
Gesell & Amatruda (1941)	Institutional: unspecified number	Birth to 2 years	No data	Gesell test	No data
Goldfarb (1945a)	Institutional: 15 Controls: 15	First test mean: 34 months Follow-up mean: 43 months	Early infancy: mean—4½ months	Intelligence tests Language test Test of motor coordination Social maturity scale Rorschach Behavior ratings	General—physical, social
Levy (1947)	Institutional: 101 Foster home controls: 129	122 under 6 months 34: 6 to 12 months, 74 over 12 months	Early infancy	Gesell test Stanford-Binet and other preschool intelligence tests Vineland Social Maturity Scale	Detailed—physical, social

TABLE 1 (*Continued*)

Investigator	Subjects	Age at time of study	Age when institutionalized	Techniques or type of data	Description of environment
Rheingold (1956)	Institutional: 16 8 controls, 8 experimental given special mothering	6–8 months	Early infancy	Cattell test Social responsiveness test	Detailed—physical, social, emotional
Skeels, Updegraff, Wellman & Williams (1938) Wellman & Pegram (1944)	Institutional: varying numbers of cases, main group 53 controls; 35 experimental, given preschool experience	1½ to 5½ years	Birth to 2 years	Intelligence test Language test Motor tests General information test Vineland Social Maturity Scale Behavior observations	Detailed—physical, social, emotional
Skeels & Dye (1938) Skeels (1942)	Institutional: 25 12 controls; 13 experimental given special stimulation	First test: controls—12 to 22 months experimental—7 to 36 months Last test: controls—5 to 9 years experimental—4½ to 9½ years	Birth to 2 years	Intelligence tests 4 follow-up tests	Detailed—physical, social, emotional
Spitz (1946)	61 infants in foundling home; 69 infants with own mothers in prison; 34 infants in own homes	Early infancy to 2½ years	Birth	Hetzer-Wolf Infant Test Clinical observation	Detailed—physical, social
Spitz & Wolf (1949)	170 infants with own mothers in prison; 61 infants in foundling home; 17 infants in own homes	Birth to 15 months	Birth	Observation Interview Hetzer-Wolf Infant Test Rorschach (mothers)	General—physical, social; detailed personality of mothers

is not a simple variable, and cannot be used as a simple research variable or explanatory concept. Even in the limited sample of institutions found in the direct studies, the environments are not identical. Qualitative as well as quantitative variations are apparent among institutions in the amount of sensory stimulation, in the consistency of mothering, in the consistency of rewards, etc.

Intellectual, Personality, and Social Characteristics Associated with Institutionalization

Despite the methodological inadequacies and the great range of antecedent conditions in the research, there is a core of consistency in the findings on the characteristics of children, adolescents, and adults with institutional backgrounds. The major characteristics associated with institutional care are: general intellectual retardation, retardation in language functions, and social and "personality" disturbances, chiefly disturbances centering around the capacity to establish and maintain close personal relationships. Within the overall consistency, however, there is significant variation. Not all children with institutional experience give evidence of intellectual or personality damage, and there is a range in the extent of injury. These variations can sometimes be related to the characteristics of the environment; sometimes significant modifying or interacting variables can be identified.

Intellectual defects. General intellectual retardation is commonly found in older children and adolescents with a history of institutionalization (Bender, 1947; Goldfarb, 1945a; Levy, 1947; Lowrey, 1940) as well as in infants and young children growing up in institutional environments (Dennis & Najarian, 1957; Fischer, 1952, 1953; Gesell & Amatruda, 1941; Skeels, Updegraff, Wellman, & Williams, 1938; Spitz, 1945, 1946). The data do not, however, permit the simple conclusion that gross intellectual deficiency is a necessary consequence of institutional experience. The incidence and degree of retardation vary considerably from one study to another. In only *some* of the studies do *some* children show severe retardation (Dennis & Najarian, 1957; Gesell & Amatruda, 1941; Goldfarb, 1945a, 1945b; Skeels et al., 1938; Spitz, 1945, 1946). In others there is only relative retardation; they are functioning on a dull-normal level (DuPan & Roth, 1955; Fischer, 1952, 1953; Freud & Burlingham, 1944; Klackenburg, 1956; Rheingold, 1956). Several factors seem to be related to the varied outcomes in intellectual functioning:

1. The amount of individualized stimulation provided in these environments seems to be significantly related to the degree of retardation. In the institutions in which attempts were made to provide individualized stimulation, and to foster a relationship between a single caretaker and infant, severe retardation was not found (DuPan & Roth, 1955; Fischer, 1952, 1953; Freud & Burlingham, 1944; Klackenburg, 1956; Rheingold, 1956).

2. The age of the child at the time of institutionalization varies greatly among the studies; several investigators have concluded that the younger the child at the time of institutionalization, the more likely is subsequent retardation (Bender, 1945, 1947; Beres & Obers, 1950; Goldfarb, 1947). The evidence is meager, consisting of data from two studies. In Goldfarb's research in which a large percentage of cases showed evidence

of retardation, the mean age of admission to the institution was 4.5 months, with only three cases over 1 year of age. Of a group of 37 adolescents and young adults studied by Beres and Obers (1950) only four were mentally retarded; all four had entered the institution under 6 months.

3. Constitutional factors. There are no direct data, but the findings that, in seemingly identical environments, some children show retardation and others do not, have been interpreted as evidence of constitutional differences in vulnerability to institutional deprivation.

4. The duration of institutionalization. The data point to a cumulative impact of the institutional environment on intellectual functioning. In most studies, with continued institutional residence, infants show a progressive drop in developmental test quotients (Dennis & Najarian, 1957; Fischer, 1952, 1953; Freud & Burlingham, 1944, no test data; Skeels, 1942; Skeels & Dye, 1939; Spitz, 1945, 1946). A few studies (DuPan & Roth, 1955; Rheingold, 1956; Rheingold & Bayley 1959) report no significant cumulative loss in intellectual functioning. Although Dennis and Najarian (1957) found a decrease in Cattell test scores in institutionalized infants between 3 and 12 months they discovered no significant retardation on the Goodenough Draw-A-Man Test among a group of older children, 4.5 to 6 years of age, who had been in the same institution for several years. They raise the interesting question as to whether an environment which fails to offer adequate intellectual stimulation to infants is necessarily retarding for preschool children.

The direct association between intellectual retardation and environmental impoverishment is dramatically emphasized by Skeels and Dye's study (1939). Retarded institutional children made significant gains in intellectual functions after special environmental stimulation. In another study (Skeels et al., 1938), the intellectual stimulation provided by an experimental nursery school in an institution was found effective in preventing deterioration in intellectual functioning. Whereas a control group showed cumulative losses in IQ scores, children given nursery school experience maintained their IQ level.

Two other studies suggest that intellectual retardation need not be attributed to some elusive, unknown aspect of the institutional environment, but can be directly related to lack of adequate stimulation. Rheingold (1943) studying infants in boarding homes found that children who shared the home with several other babies had significantly lower developmental test scores than infants who were "only children" in the boarding homes. Coleman and Provence (1957) observed retardation similar to the institutional pattern in children living in very unstimulating home environments.

Analysis of the separate aspects of intellectual functioning indicates that all functions are not equally affected by institutional living. Consistent evidence of retardation is found in language, in time and space concepts, and in capacity for abstract conceptualization.

Language is one function in which severe retardation has been found repeatedly in institutionalized infants and young children (Brodbeck & Irwin, 1946; DuPan & Roth, 1955; Fischer, 1952, 1953; Freud & Burlingham, 1944; Gesell & Amatruda, 1941; Skeels et al., 1938; Rheingold &

Bayley, 1959) as well as in older children and adults with an institutional history (Bender, 1945, 1947; Goldfarb, 1945a; Haggerty, 1959; Lowery, 1940). There is disagreement in the literature on institutionalization only in the age at which language functions first seem to be affected. Brodbeck and Irwin (1946) found evidence of retardation in institutionalized infants in the first few months of life, whereas Freud and Burlingham (1944) report no indications of language retardation before 12 months. Brodbeck and Irwin's data were based on careful phonetic analysis of speech sounds, whereas Freud and Burlingham had no systematic language data on infants.

With regard to the etiology of language retardation, Fischer (1952, 1953) notes that in many institutions there is little reinforcement by adults of the infant's vocalizations, and consequently reduced opportunity for the child to acquire the signal functions and expressive functions of language. Recent data on the conditioning of vocalizations in infants (Rheingold, Gewirtz, & Ross, 1959) give evidence of the role of reinforcement in young infant's vocalizations. Early studies of language development (Day, 1932; Van Alstyne, 1929) pointed to a direct relationship between amount of environmental stimulation (e.g., number of hours the child was read to, "extensions of the environment") and vocabulary and sentence length in preschool children. On the simplest level, language retardation, like general intellectual retardation, can be related to inadequate language stimulation. Lack of motivation for imitative behavior may interact with inadequate reinforcement of speech sounds in determining language retardation.

Serious *defects in time and spatial concepts* in older children have been reported in clinical descriptions by Goldfarb (1945a, 1949) and Bender (1945, 1947). Poor memory for past events is linked by Bender with such character defects as inability to benefit from past mistakes, lack of future goals, and weak motivation to control behavior for future gains. Goldfarb relates social maladjustment to difficulties in time and spatial concepts. As a result of these conceptual difficulties, disregard of school and family rules occurs.

Disturbances in abstract thinking were also found by Bender (1947) and Goldfarb (1943b) in school aged children and in adolescents with an institutional background. Goldfarb (1945b) describes as characteristic of these children "an unusually defective level of conceptualization . . . manifested in difficulty in organizing a variety of stimuli meaningfully and in abstracting relationships" (p. 251). On the Rorschach test, adolescents with an institutional background showed "an unusual adherence to a concrete attitude and inadequate conceptualization" (1943a, p. 222).

Motor functions. Motor development seems to be less significantly affected than any other aspect of development, although there are markedly discrepant reports. DuPan and Roth (1955) and Fischer (1953) conclude that there is no significant retardation in motor development during the first year among institutionalized children, Freud and Burlingham (1944) report accelerated development during the early part of the second year, while Spitz (1946) notes marked retardation in motor functions during the first and second years. Differing opportunities for the exercise of developing motor functions in different institutional

TABLE 2

RESEARCH ON INSTITUTIONALIZATION: RETROSPECTIVE STUDIES

Investigator	Subjects	Age at time of study	Age when institutionalized	Duration of institutionalization	Techniques or type of data	Data on early experiences	Contaminating conditions
Bender (1947)	5000 clinic cases	Preadolescence	Birth to middle childhood	Range; not specified	Case history Psychiatric diagnosis	General retrospective	Repeated separations Rejection
Bender & Yarnell (1941)	250 clinic cases	1–6 years	Birth to 6 years	Range; not specified	Intelligence tests Psychiatric diagnosis Case history	General retrospective	
Beres & Obers (1950)	37 clinic cases with institutional background	Adolescence and adulthood	Birth to 12 months	Varying periods up to 4 years Average: 3 years	Case history Psychiatric diagnosis Intelligence test	Detailed case history	Separation and rejection
Bodman et al. (1950)	51 cases with institutional background 52 controls	Early adolescence	16 cases under 2 years Average: 4.4 years	Range from 3 to 15 years Average: 9.6 years	Vineland Social Maturity Scale Case history	General data on variety of institutions	High incidence of mentally defective or disturbed parents Several changes in institutions
Goldfarb (1943b)	20 children with institutional background 20 foster home controls	6–10 years Follow-up	1 to 24 months	3 years	Baruch Preschool Checklist Newell Problem Checklist	General retrospective data	Repeated separations from foster mothers
Goldfarb (1945b)	15 children with institutional background 15 foster home controls	Mean: 12 years	Early infancy	2½ to 3 years	Intelligence tests Concept formation tests—Weigl, Goldstein-Scheerer Clinical assessment of personal-social functioning	Some retrospective data	Repeated separations from foster mothers
Goldfarb (1944)	40 children with institutional background 40 foster home controls	Mean: 7½ years Follow-up	Early infancy	Average: 34 months	Analysis of problems and reasons for replacement	Some retrospective data	Repeated separations from foster mothers Maternal rejection
Goldfarb (1947)	15 well-adjusted and 15 poorly adjusted children with institutional background	Mean: 14½ years	Poorly adjusted mean: 5.8 months Well-adjusted mean: 10.9 months	Poorly adjusted: 34 months Well-adjusted: 25 months	Caseworker's ratings on adjustment	Detailed retrospective data	Repeated separations
Goldfarb (1949)	15 institutional 15 schizophrenic 15 foster home	Mean: 12 years	Early infancy mean: 4½ months	Group average: 39 months	Rorschach test Intelligence test	No data	Not reported
Haggerty (1959)	100 social agency cases with institutional background	Mean age: 12.7 years	"First few years of life"	Average: 3½ years	Analysis of language samples	No data	Separation
Lowrey (1940)	28 psychiatric clinic cases with institutional background	Range from 3 to 6 years	Range from 2 weeks to 34 months	Range from 6 to 42 months	Case history Intelligence tests	Variable retrospective history	Repeated separations

settings may be involved (Skeels et al., 1938).

Both extremes in *activity level* are found in institutionalized infants. Hyperactivity is sometimes noted (Fischer, 1952) but more common is a lowered activity level, associated with the general passivity noted as part of the pattern of intellectual retardation. There are only vague indications in the data of some factors which may account for these different findings: constitutional differences among infants, the age or developmental level at the time of institutionalization, and the length of institutionalization. For instance, in the initial stages of institutionalization, hyperactivity is often found, with lowered activity level more common after prolonged institutional residence.

Motor disturbances in the form of bizarre stereotyped motor patterns suggestive of neurological damage have been reported by Spitz (1946) in infants after a long period of institutional residence; similar but less extreme motor disturbances were noted by Fischer (1952, 1953). In older children, Bender (1947) and Goldfarb (1943a, 1945b, 1947) found hyperkinetic behavior, a pattern considered part of a syndrome of impulsivity, with psychogenic rather than neurogenic bases.

The findings on deviant motor patterns and the data on defects in conceptual thinking suggest the possibility of central nervous system damage as a result of institutionalization. The evidence is not very strong, however, nor are there clear bases in these data for hypothesizing the conditions under which irreversible neurological damage might occur.

Social and personality disturbances. Although the institutional syndrome has most frequently been described in terms of social and personality disturbances, in many respects the data are less clear than are the findings on intellectual development. Personality data are based primarily on clinical impressions, and the characteristics described are usually at the extreme end of the scale, reflecting exaggerated pathology or a complete lack of capacity, rather than a relative deficiency.

Interpersonal relationships. The major deviations reported in the literature are in the area of interpersonal relationship. Two overtly dissimilar, but dynamically related, types of interpersonal disturbance have been described: social apathy manifested by indifference to social attachments, and "affect hunger" characterized by incessant and insatiable seeking of affection. Several retrospective studies report a syndrome in older children and adolescents described as an inability to establish close, warm personal relationships (Bender, 1947; Bender & Yarnell, 1941; Goldfarb, 1943a, 1945b, 1949; Lowrey, 1940), a personality pattern labeled the "affectionless character" by Bowlby (1944), and one which Bender (1947) identifies as a psychopathic behavior disorder.

In the contemporaneous studies of infants in institutions, social apathy is described in terms of several specific response patterns:

1. Inadequate social responsiveness, as evidenced by a complete lack of social initiative, by withdrawn or apathetic response to social approaches (Bakwin, 1949; Fischer, 1952, 1953; Freud & Burlingham, 1944), or in depressed scores on the social sector of developmental tests (DuPan & Roth, 1955; Fischer, 1952, 1953)

2. An indifference to social attachments, manifested by lack of any

significant attachments or meaningful relationships with caretakers in the institution (Freud & Burlingham, 1944; Rheingold, 1956)

3. Inadequate social discrimination as evidenced by failure to give differentiated responses to strangers and familiar caretakers (Freud & Burlingham, 1944)

4. A lack of normal social sensitivity, indicated by inability to respond discriminatively to different kinds of emotional expression (Freud & Burlingham, 1944)

The specificity of the relationship between social stimulation and social responsiveness in infancy is pointed up by Rheingold's data (1956). Infants in an institution who were given intensive social stimulation by one mother-figure, from the sixth to the eighth month of life, showed significantly greater social responsiveness than control subjects cared for by the more usual institutional routine. General developmental progress was not affected, however, by this special type of stimulation. In a follow-up of these children in adoptive homes at 19 months of age, Rheingold and Bayley (1959) found no evidence of any lasting impact of this special experience.

The syndrome of "affect hunger" characterized by indiscriminate and insatiable demands for attention and affection is less common than social apathy. It is reported in several retrospective studies (Bender, 1945, 1947; Goldfarb, 1945b; Lowrey, 1940), but in only one contemporaneous study (Freud & Burlingham, 1944), in which children in an institution are described as "exacting, demanding, apparently passionate, but always disappointed in new attachments" (p. 58). A similar, but less intense pattern of indiscriminate sociability among 6–8 month old infants was observed by Rheingold (1956). Freud and Burlingham also noted in infants an associated pattern of exhibitionism, involving indiscriminate display of themselves before strangers.

Behavioral deviations considered symptomatic of disturbances in ego and superego development have been reported in older children (Bender, 1945, 1947; Beres & Obers, 1950; Goldfarb, 1943a, 1949; Lowrey, 1940). Frequently noted is a pattern of diffuse and impulsive behavior suggesting a lack of normal inhibitory controls. In these children overt antisocial and aggressive behavior is often found. Bender and Goldfarb both note a lack of normal anxiety or guilt about aggression, a low frustration tolerance, a lack of goal-directedness, and low achievement motivation. Goldfarb (1943a) summarizes the personality pattern as impoverished, meager, and undifferentiated, deficient in inhibition and control. Even as late as adolescence, the institution children show the simple, unrefined, undifferentiated kind of behavior typical of preschool children.

Beres and Obers (1950) is the one psychiatrically oriented study which raises some question as to the extent of personality damage resulting from institutionalization. They note a similar underlying pathology in all cases—a distortion in psychic structure, an immature ego, and deficient superego development—but conclude that by late adolescence about half of their 37 cases were making a favorable overt adjustment. They were

> functioning well, whether in work situation or at school . . . and presented no evidence of overt disturbance in their behavior or in their relationships within their families or among friends (p. 228).

This study points up the problem for research of making a valid distinction between mental health and pathology. These conclusions illustrate sharply the conflict between a definition of mental health based on overt behavior and a definition derived from a psychodynamic assessment of strengths and liabilities.

In looking to the direct studies for clues to the antecedents of personality deviations in older children, one is disappointed by the limited data on the personality characteristics of infants in institutions. The meager data on infants suggest some precursors of defective ego and superego development such as failure to show imitative behavior at the appropriate developmental period (Freud & Burlingham, 1944; Fischer, 1953). The conflicting findings on autoerotic activity emphasize the lack of agreement as to what constitutes normal behavior in infancy. Freud and Burlingham (1944) as well as Fischer (1952, 1953) describe a high incidence of thumbsucking, rocking, head-banging in young infants, and masturbation in older children. Spitz and Wolf (1949), on the other hand, found "practically no autoerotic activities" among the infants in the foundling home. They hypothesize that an emotional relationship between the child and a mother-figure is a prerequisite for the appearance of autoerotic activities.

Few direct studies give information on the age at which personality disturbances first become evident. In most of this research, the youngest children are over 6 months at the time of study. Where younger children have been studied, frequently no data are given on social or personality characteristics. Only two studies offer data on the age at which personality disturbances are first noted. Freud and Burlingham (1944) note that infants in their institution did not show signs of social retardation before 5 months. Gesell and Amatruda (1941) report first signs of "social ineptness" evident at 24 weeks.

The one experimental study on human infants (Dennis, 1941) is often cited as evidence that early sensory and social deprivation need have no impact on development. Dennis found no significant retardation in a pair of twins who were given "minimum" social and sensory stimulation during the first 7 months of life. Stone (1954) on the basis of a careful analysis of a later report (Dennis & Dennis, 1951) suggests that minimum stimulation probably represented minimal adequate stimulation, much more than that provided in many institutional environments. In Dennis' study the infants were handled for the normal routines, and there was a consistent mother person. The fact that these conditions did not continue much beyond the first half year may also be significant.

Many ad hoc theories have been offered to account for the intellectual and language retardation, the specific defects in abstract thinking, and the varied social and personality disturbances associated with institutionalization. The explanations which offer "maternal deprivation" as the basic etiological entity tend, on the whole, to be vague and generalized, and offer little basis for systematic research. With regard to abstract thought, Bender (1946) states:

> The earliest identification with the mother and her continuous affectional care is necessary during the period of habit training and the rapid development of language and the formation of concepts within the family unit. Otherwise the higher semantic and social development and the expansion of the educational capacities does not take place (p. 76).

(Quoted by permission of Child Study Association of America.)

Regarding time concepts, she speculates, "It appears that we develop a concept of time in the passage of time in our early love relationships with our mother" (p. 96). Kardiner (1954) suggests that the sense of time develops in relation to the child's activities in looking forward to gratification. Goldfarb (1955) hypothesizes that lack of an adult identification model (in institutions) inhibits the development of functions such as language, which are dependent on social forms of imitation and communication. Impairment in abstract thinking is interpreted (Goldfarb, 1955) in terms of Stern's theory (1938) which postulates that the development of conceptual thinking is dependent on the growth of a sense of continuity of the self. According to Stern, the grasp of identity, as well as judgments of equality, similarity, and difference are all derived from the sense of continuity of self. At first these judgments are related to concrete personal events; eventually, they are separated from them and become abstract. Without continuity of mothering in an institution, Goldfarb contends the normal development of the self-concept is impaired, with resulting defects in abstract thought processes. Social and personality disturbances are linked directly to lack of opportunity for close human relationships in infancy in institutional environments. Goldfarb attributes defective ego and superego development to inadequate opportunity for the child to identify with parental figures and to internalize the parental image. Bender (1946) describes the etiology of personality disturbances in similar terms:

> There is a primary defect in ability to identify in their relationships with other people . . . due to the fact that they never experienced a continuous identification during the infantile period from the early weeks through the period when language and social concepts of right and wrong are normally built up and when psychosexual and personality development are proceeding (p. 76). (Quoted by permission of Child Study Association of America).

She hypothesizes that anxiety and guilt arise in reaction to "threats to object relationship or identification processes" (p. 76). Lack of anxiety and inability to feel guilt are related to the lack of capacity to identify or form object relationships.

Analysis of environmental variables in the research literature points to some more discrete factors than maternal deprivation in the institutional setting. This elusive variable, maternal deprivation, can be analyzed in terms of variables more amenable to research, e.g., amount and quality of tactile, auditory, or visual stimulation; reinforcement schedules; etc. Harlow's (1958) research on infant primates has demonstrated the efficacy for research of analyzing mothering in terms of simple stimulus conditions, such as contact stimulation. The discrepancies in the findings of the research on institutionalization suggest the need to consider interacting variables, such as constitutional differences in vulnerability, varying sensitivities at different developmental stages, etc., in formulating hypotheses for more critical research testing.

Maternal Separation

Maternal separation has never been studied under pure conditions. Most often separation has been associated with other traumatic events such as illness and hospitalization or operative procedures, and often with parental rejection or death or disability of a parent. Frequently sepa-

TABLE 3

RESEARCH ON MATERNAL SEPARATION

Investigator	Subjects	Age at time of study	Age at time of experience	Techniques or type of data	Data on early experiences	Contaminating conditions
Ainsworth & Boston (1952)	One case	Observation: 3 years Follow-up tests: 5 to 6½ years	13 months	Rorschach CAT Stanford-Binet Weigl-Goldstein Sorting Test Goldstein-Scheerer Cube Test	Retrospective report	Hospitalization for tuberculosis
Berg & Cohen (1959)	40 schizophrenic women in mental hospital 40 neurotic women	20–40 years	Birth to adulthood	Case history	Limited retrospective data	Rejection
Bowlby (1944)	44 juvenile thieves	5.7 to 17 years	Birth to adolescence	Case history Psychiatric diagnosis	Variable retrospective data	Institutionalization Rejection
Bowlby (1953b)	49 children in residential nurseries or hospitals	12–48 months	12–24 months	Clinical observation	Direct observation	Institutionalization Rejection Hospitalization for illness
Bowlby, Ainsworth, Boston, & Rosenbluth (1956)	60 children with previous sanitarium experience 57 controls	6–14 years	Range: Birth to 4 years	Intelligence test Clinical evaluation by teacher, psychologist, psychiatrist, social worker	General retrospective data	Rejection Hospitalization for tuberculosis
Edelston (1943)	42 children hospitalized for illness	2½–15 years	Range from early infancy	Clinical observation	Limited retrospective data	Illness Rejection
Heinicke (1956)	Children in residential and day nurseries	15–31 months	12–30 months	Standardized observation and doll play	Direct observation	None reported
Lewis (1954)	500 children in reception center	Under 5 to over 15 years	Birth to adolescence	Clinical assessment	Variable retrospective data	Institutionalization Rejection
Robertson & Bowlby (1952)	Unspecified number of children in hospitals	18–24 months	18–24 months	Clinical observation	Direct observation	Hospitalization
Roundinesco, David, & Nicolas (1952)	20 children placed in institution	12–17 months	12–17 months	Clinical observation	Direct observation	Institutionalization
Schaffer (1958)	76 infants in hospital for illness	3–51 weeks	3–51 weeks	Cattell Infant Test Standardized observation Home follow-up	Direct observation	Illness Hospitalization
Spitz & Wolf (1946)	123 children in a nursery	14 days to 18 months	5–7 months	Clinical observation	Direct observation	None reported

ration from the parents has been followed by institutional placement with the result that the impact of institutional influences is superimposed on the loss of parental figures. In the literature on separation, the role of such contaminating variables has not been distinguished from the effects of a break in continuity of relationship with the mother. Spitz and Wolf's (1946) is the only study in which the physical environment remained unchanged following separation; it is one of the few studies in which the quality of the mother-child relationship prior to separation had been studied.

Most of the research is contemporaneous, reporting on the reactions of children at the time of separation. The long-term effects are almost unknown. Follow-up data more than a year later are given in a few studies (Bowlby, Ainsworth, Boston, & Rosenbluth, 1956; Lewis, 1954; Spitz, 1954a, 1954b; Spitz & Wolf, 1946), but in these studies there are many contaminating conditions, e.g., severely disturbed parental relationships, repeated separations, intermittent institutionalization.

Immediate Reactions to Separation

Despite the many different conditions associated with the separation experience, there is some degree of consistency in the findings reported on immediate and short-term reactions of infants and preschool children to separation. In each of the studies some children develop apparently severe reactions, and the behavior sequences in these extreme cases appear to be dynamically similar (Bowlby, 1953b; Robertson & Bowlby, 1952; Roundinesco, David, & Nicolas, 1952; Spitz & Wolf, 1946). The characteristic sequence of responses begins with active protest and violent emotional reactions, such as intense and prolonged crying and active reaching out to people, in apparent attempts to bring back the mother or to find a substitute. In time this behavior is followed by active rejection of adults, and finally by apathy and withdrawal of interest in people, accompanied by a decrease in general activity level. Robertson and Bowlby characterize this latter phase as "mourning"; Spitz and Wolf label it "anaclitic depression." Feeding disturbances—refusal of food, sometimes pathological appetite—and regression in motor and other functions are also reported. When the mother is not restored, Spitz found symptoms of progressive deterioration in infants, a complete withdrawal from social interaction, a sharp drop in developmental level on infant tests, and extreme physical debilitation, with loss of weight and increased susceptibility to infections. In older children (over 12 months) marked physical and intellectual deterioration have not been reported, but severe disturbances in interpersonal relationships have been noted (Bowlby, 1953b; Robertson & Bowlby, 1952). The "mourning phase" in infants and young children is followed by behavior described as a "denial of the need for his own mother," which Robertson and Bowlby interpret as an indication of a repression of the mother image. The child shows no apparent recognition of his own mother, but may transfer his attachment to a substitute mother. (There has been some controversy as to whether such behavior can be interpreted as evidence of repression or whether it should be considered more simply as a denial mechanism—Bowlby, 1953a; Heinicke, 1956.) If no substitute mother is available, the child may show promiscuously friendly behavior, using adults in an instrumental way, but without establishing meaningful attachments. Such behavior Bowlby considers indicative of a repression of all need for mothering, the prelude to a psychopathic character development. If, however, the child is reunited with his mother before the need for mothering is completely repressed (after some unstated critical time interval) the behavior pattern is believed to be reversible. The child is able on return to his mother to reestablish a relationship with her, although there may be several months of difficult adjustment, with irritability, impulsive expression of feelings, and an exaggeratedly intense attachment.

These descriptions of the reactions of young children to conditions involving loss of a mother-figure have provided the basis for most of the generalizations about the severe effects of maternal separation. The dramatic character of these changes has overshadowed the significant fact that a substantial portion of the children in each study did not show severe reactions to separation. In Spitz's study of 123 infants separated from their mothers between 6 and 8 months of age, severe reactions occurred in only 19 cases. Although in Robertson and Bowlby's (1952) research on 45 children ranging in age from 4 months to 4 years, all but three are reported to have shown some reaction; the intensity and duration of the reactions are not clearly specified. Less than half, 20 cases, are reported as showing "acute fretting," a behavior pattern which is not well-defined. The reported duration of the reaction varied from 1 to 17 days. There are no data on the number of children who showed prolonged reactions.

In a careful study of the reactions to hospitalization of 76 infants under 1 year of age (ranging from 3 to 51 weeks) Schaffer (1958) found that reactions varied with age. Infants over 7 months of age showed overt social and emotional reactions, such as excessive crying, fear of strangers, clinging and overdependence on the mother. Infants under 7 months evidenced more global disturbances, i.e., somatic upsets, blank facial expression, extreme preoccupation with the environment. Schaffer relates the global disturbances to sensory deprivation, whereas the social disturbance at the later age, an age at which more differentiated relationship with the mother exists, are interpreted as reactions to separation from the mother.

Heinicke's research (1956) points to less severe effects of simpler, less complicated separation situations. He found no extreme behavioral disturbances in two groups of children, 15 to 30 months of age, with different separation experiences, one group in a residential nursery, the other in a day nursery. The children in the residential nursery did show more overt and more intense aggression, greater frequency of autoerotic activities, and more frequent lapses in sphincter control. These findings are interpreted as indicating an imbalance between the child's impulses and his power to control and organize these impulses in relation to the external world.

Long-Term Effects of Separation

Conclusions about the long-term effects of separation are very tenuous. They are based on a few studies in which the information about the early history is not well-documented.

In an earlier study of 44 juvenile thieves, Bowlby (1944) concluded that separation experiences in childhood resulted in a character disorder distinguished by a "lack of affection or feeling for anyone." The conclusions are based on clinical findings that 12 out of 14 cases diagnosed as "affectionless characters" had been separated from their mothers in infancy or early childhood. Some of these children had been hospitalized for illness without any contact with their mothers over a long period of time, others had experienced frequent changes in foster mothers, and some had been institutionalized for long periods during infancy.

In a follow-up study of 60 children between 6 and 13 years of age, who had been in a sanitarium for tuberculosis for varying periods of time before their fourth birthday, Bowlby et al. (1956) found less serious long-term effects than in the earlier studies. No statistically significant difference in intelligence was found between the control and the sanitarium group. In personality characteristics, the sanitarium children were judged as showing tendencies towards withdrawal and apathy, as well as greater aggressiveness. On the basis of the psychiatric social worker's interview with the parents, 63% of the children were rated as maladjusted, 13% were considered well-adjusted, and 21% adjusted but with minor problems. Bowlby et al. conclude that "outcome is immensely varied, and of those who are damaged, only a small minority develop those vary serious disabilities of personality which first drew attention to the pathogenic nature of the experience" (p. 240). They suggest that the potentially damaging effects of separation should not be minimized, but concede that "some of the workers who first drew attention to the dangers of maternal deprivation resulting from separation have tended on occasion to overstate their case" (p. 242).

The findings of Lewis (1954) are sometimes cited as evidence that early separation need not necessarily have lasting harmful effects. Among a group of 500 children who were studied in a reception center shortly after being separated from their parents, only 19 showed "morbid lack of affective responsiveness" (p. 41). Follow-up data were obtained on 240 of these children, 2–3.5 years later. Only 100 had a personal follow-up by a psychiatric social worker and a psychiatrist; information on the others was obtained through letters from social workers who had some contact with the children. Of the 100 more intensively studied children, only three were diagnosed as having marked personality disorders, 22 were having some difficulties in relationships, and 36 were showing mild neurotic symptoms or mild delinquent behavior. With reference to the timing of separation, Lewis concludes that "separation from the mother before the age of five years was a prognostically adverse feature" (p. 122). Apparently this is a clinically based conclusion, since the data presented in the tables show no significant differences between the children separated before 5 years of age and those separated after 5.

Data from several studies indicate that the impact of separation is modified by the character of the mother-child relationship preceding the separation experience and the adequacy of the substitute mothering following separation. Spitz and Wolf (1946) noted that the infants who did not develop severe depressive reactions were those separated from "poor mothers," and conclude that the better the mother-child relationship preceding separation, the

TABLE 4

Research on Multiple Mothering

Investigators	Subjects	Techniques	Age at time of experience	Age at study	Data on early experiences
Rabin (1957)	38 children from kibbutz and 34 controls from neighboring villages	Rorschach	Birth to time of study	9–11 years	General description of environment
Rabin (1958a)	24 infants and 40 children in kibbutz 20 control infants and 40 control children	Rorschach Vineland Social Maturity Goodenough Draw-A-Man Griffiths Infant Scale	Birth to time of study	9–17 months 9–11 years	General description of environment

more severe the immediate reactions. Lewis (1954), on the other hand, found a higher proportion of children who had been separated from normally affectionate mothers in "good" or "fair" condition than those who had not received "adequate" affection. It might be hypothesized that a close relationship with a mother-figure preceding separation will be followed by more severe immediate reaction but will be ultimately more favorable than a poor antecedent relationship. Children who have experienced a close relationship in infancy may be better prepared to form new attachments in later life than children without any experience of close relationships.

The amount, the quality, and the consistency of substitute mothering will presumably influence the intensity of immediate reactions as well as the long-term personality consequences. Spitz and Wolf (1946) concluded that infants who were provided with a satisfactory substitute mother did not develop the depressive syndrome. (There were no independent criteria of the adequacy of substitute mothering. The substitute relationship was considered satisfactory in those cases which did not develop depressive symptoms.) Robertson and Bowlby (1952) also note that where an adequate substitute mother was provided, there was not a complete withdrawal from social contact.

Multiple Mothering

Serious personality difficulties in later life have been postulated as a consequence of multiple mothering in infancy and early childhood. There has been little research, and in most of the clinical observations multiple mothering has been associated with impersonal or rejecting maternal care. The underlying assumption in much of the literature is that inadequate maternal care is a necessary concomitant of situations in which there is more than one mother-figure. Multiple mothering has never been very precisely defined. In its most general sense, it refers to an environmental setting in which a number of different persons perform the maternal functions for the child, with varying degrees of adequacy and with varying degrees of consistency. From the child's viewpoint, it may mean that there is no single person to whom he can relate as a major source of gratification and on whom his dependency needs can be focused. In some situations the biological mother may share the mothering functions with other chosen women; in other circumstances no biological tie exists between the child and the several mothers. Some current studies in home management houses, a few reports on the Israeli kibbutzim, and a very few anthropological reports provide all the available data on the effects of multiple mothering.

In the anthropological accounts of multiple mothering in different cultural contexts (DuBois, 1944; Eggan, 1945; Mead, 1935; Roscoe, 1953) there are variations in the number of people who share mothering functions as well as variations in the role of the natural mother. In cultures in which the extended family is the traditional pattern, the mothering functions may be shared by the mother, grandmother, aunts, and other female relatives of the child; in some groups, male relatives may take over some maternal functions. The biological mother may be clearly identified as the central, most significant person in some cultures; in others she may be assigned a very secondary role.

In Western cultures, grandmothers frequently assume some of the mothering functions, and in some social groups, child nurses play an important role. In the pre-Civil War Southern plantation class group, many mothering functions were taken over by the Negro nurse. The line of demarcation between supplemental maternal care and multiple mothering has never been very clear.

In none of these situations are disturbances in infant functioning associated with multiple mothering practices, nor are later personality characteristics or deviations attributed to this aspect of early maternal care.

The Israeli kibbutzim provide an unique set of conditions of multiple mothering. In this setting, there are two mother-figures, the natural mother and the metapelet, the children's caretaker, each of whom has very distinctive functions. The major share of the daily routine care as well as major training functions, such as toileting and impulse control, are assumed by the caretaker in the communal nurseries. The mother's contacts with the child tend to be limited to scheduled periods during the day, which are free periods and do not involve traditional family routines. The mother seems to function solely as an agent to provide affectional gratification, although obviously the extent of the mother's influence, as well as the specific areas of influence on the child's development, will vary with her concept of her role and with her personality characteristics.

There are several impressionistic reports (Golan, 1958; Irvine, 1952; Rapaport, 1958) and a few systematic studies (Rabin, 1957, 1958a) of the development of infants and children in the Israeli kibbutzim. Rabin (1958a), using the Griffiths Infant

Developmental Scale, found slight developmental retardation in infants between 9 and 17 months of age living in a communal nursery. In only one sector of development—the personal-social area—were these infants significantly retarded. Rabin attributes this retardation to less individual stimulation in the kibbutzim as compared to a normal home environment. This study represents the only reported research in a setting in which there may be deprivation in the amount of stimulation without concomitant lack of affectional interchange with the mother.

In an attempt to assess the long-term effects of living under these special conditions of maternal care in the kibbutz, Rabin (1958a) studied a group of children, between 9 and 11 years of age, who had lived in this environment from infancy. He found no evidences of retardation (using the Goodenough Draw-A-Man Test), nor were there any indications of personality distortions. On the contrary, Rorschach data are interpreted as indicating that the children from the communal settlements showed "better emotional control and greater overall maturity." In ego-strength (using Beck's index) they were judged superior to the control group of children living with their parents. Rabin interprets these findings as evidence of the important role of later experiences in personality development.

In another study, Rabin (1958b) compared the psychosexual development of 10-year-old kibbutzim reared boys with boys from patriarchal type families. Using the Blacky test, he found significant differences, consistent with theoretical expectations. The kibbutz boys showed less "oedipal intensity," more diffuse positive identification with their fathers, and less intense sibling rivalry. This study also points up the fact that multiple mothering is only one of the significant factors which differentiate the kibbutz from the "normal" family setting. As in the case with other conditions associated with maternal deprivation the kibbutz is atypical in regard to the absence of the father.

Home management houses provide a setting in which multiple mothering occurs without associated deprivation of social stimulation. These houses are set up in university home economics departments to provide practical experience in child care for the students. The infant is separated from his foster mother or removed from a familiar institutional environment and placed in the home management house for a period of several weeks to several months. He is cared for by a number of young women, each of whom assumes primary responsibility for mothering activities for a limited period of time, usually about one week. There is one continuous figure—the instructor in the house—with whom the infant can maintain a relationship; she assumes some of the ordinary child care functions. In the course of his residence in the home management house, the infant may have 15 to 20 different "mothers." In this setting he receives much attention and stimulation from many different "mother-figures." Following his residence in the home management house, the infant is usually placed in a foster or adoptive home. The follow-up studies and the several direct studies of children in home management houses (Gardner, Pease, & Hawkes, 1959; Gardner & Swiger, 1958) are in agreement in finding no evidence of intellectual retardation and no gross personality disturbances. The long-term effects have not yet been evaluated.

These three settings—the home management house, the kibbutz, and the extended family—are comparable in only one respect; the mothering functions are distributed among several different persons. They differ in regard to the continuity of the mother-figure, in the role played by the substitute mothers, and in the amount of social stimulation given to infant. In some situations, because of the high adult-child ratio, it is likely that the infant will receive more sensory as well as more social stimulation than the child in an average family home. For infants, the kibbutz may be similar to an institutional setting in terms of the amount of individual social stimulation provided. It is clear that none of these conditions necessarily involves severe deprivation of mothering, but the mothering experience of children in these settings may differ significantly from that of children in homes with one mother-figure.

None of these studies provides a crucial test of the prevalent hypothesis that multiple mothering results in a diffusion of the mother-image. This theory, developed in the context of institutional care, holds that the child who is cared for by a number of different persons cannot develop a focused image of one significant mother-person in infancy, and consequently, will have difficulties in relationships in later life. On the whole, the few relevant pieces of research suggest that multiple mothering per se is not necessarily damaging to the child.

Distortions in the Mother-Child Relationship

Although distortions in the mother-child relationship have frequently been included in the concept of maternal deprivation, in this report we shall not attempt any comprehensive review of this vast clinical literature. Institutionalization, separation, and multiple mothering represent deviations from a cultural norm of "mothering" primarily on the dimension of amount or consistency of contact with the mother. Under the category of distortions in the mother-child relationship are subsumed all the deviations in maternal relationships which usually have as their antecedents disturbances in the character or personality of the mother. These disturbances in maternal relationships are manifested in overtly or covertly hostile or rejecting behavior, sometimes more subtly in overprotective behavior, and often in unpredictable swings from affection to rejection or in ambivalent behavior. As distinguished from a lack of social stimulation, a lack of responsiveness, and the lack of a mother-figure, this type of deviation in maternal care tends to be characterized by either very strong emotional stimulation, or by stimulation with a preponderance of negative affect. In contrast to institutional care, there may even be very intense intellectual stimulation.

The literature on distorted maternal relationships suggests a somewhat different kind of personality outcome from the psychopathic or affectionless character. The personality distortions tend to be in the schizophrenic, depressive, and neurotic categories. Again there may be rather specific antecedent conditions and organismic vulnerabilities associated with these types of personality deviations (Spitz, 1951). A critical review pointed towards a clarification of the variables and an analysis of the many ad hoc theories concerning distorted mother-child relationships is very much needed.

Some Theoretical Issues and Research Implications

The data from the research on institutionalization, maternal separation, and multiple mothering have relevance for a number of fundamental issues in developmental theory: questions concerning the kinds of environmental conditions which facilitate, inhibit, or distort normal developmental progress; the conditions which influence the reversibility of effects of events in infancy and early childhood; and the extent to which the timing of an experience, i.e., the developmental stage at which it occurs, determines its specific impact.

In theories of the effects of early infantile experiences on later development, two concepts have been prominent: deprivation and stress. Although all the intricacies of the mother-child relationship cannot be conceptualized adequately in terms of these concepts, some of the environmental conditions and events found in the research on maternal deprivation can be ordered meaningfully in these terms. Deprivation is a key concept in the analysis of institutional environments. Many of the circumstances associated with maternal separation and multiple mothering can be ordered in terms of the concept of stress.

Deprivation

In institutional settings several types of deprivation, each with potentially different developmental implications, can be distinguished: sensory deprivation, social deprivation, and emotional deprivation. In many settings all three types of deprivation occur and are complexly interrelated, but they do not necessarily vary concomitantly, and they can be independently manipulated in research.

The studies on sensory deprivation in animals indicate that complete restriction of perceptual experience in early life results in permanent impairment in the functions in which deprivation occurs. In the most extreme institutional environments the degree of sensory deprivation is less severe than in the animal studies. Nevertheless, developmental retardation is found, with the extent of retardation corresponding to the degree of sensory deprivation.

Social deprivation probably acts in a similar way as deprivation of sensory stimulation, leading to disturbances in social functioning, such as, social apathy and social hyperresponsiveness. The simplest hypothesis relates social apathy to inadequate social stimulation during a developmental period which is critical for the acquisition of social responsiveness. If social deprivation occurs after appropriate social responses have been learned, affect hunger or intensified seeking of social response may occur. Although social deprivation is less amenable to experimental manipulation than is sensory deprivation, in natural situations, some simple indices can be used, such as the number of persons with whom the infant has contact during a 24-hour period, the amount of time during which he receives stimulation.

Emotional deprivation has been used popularly and in clinical writings as a catchall term to include deprivation of social, sensory, and affectional stimulation. For research, a more precise usage in terms of deprivation of affective stimulation may be useful. The term, emotional deprivation, can be restricted to characterize an environment with neutral feeling tone or without variation in feeling tone, an environment similar in some respects to the monotonous, bland environment described under sensory deprivation. Emotional apathy, withdrawn behavior, lack of differentiation of affect, and insensitivity to feelings or emotional nuances in others are characteristics which might be related to early emotional deprivation. Within this concept of emotional deprivation, simple objective measures are also possible, e.g., ratings of intensity of positive or negative affect, amount of time during a 24-hour period in which different types and intensities of affective stimulation are provided.

In addition to independent manipulation of each of these types of stimulation—sensory, social, and emotional—in more focused research there might be systematic variation in several dimensions of stimulation: quality of stimulation, e.g., monotonous, varied; intensity; frequency; regularity; cumulative duration of deprivation; sensory modalities in which deprivation occurs.

Stress Consequent to Change

Critical research on maternal separation requires a distinction between the event of separation and later conditions often associated with separation which may be similar to those described under deprivation. The event of separation is associated with significant changes in the physical, and social environments, changes which may be stressful for the young child. In the physical environment, the changes involve the disappearance of familiar objects, sounds, smells, and tactile stimuli; in the social environment, there may be changes in the amount and quality of social stimulation. The new environment may provide more tactile stimulation and less verbal stimulation. There may be modifications in the speed as well as kind of response to the child, e.g., the new caretaker may ignore the child's crying, or she may reward it by tactile stimulation rather than by oral gratification. For the infant or young child, these changes result in a loss of environmental predictability. The degree of stress experienced is likely to vary with the degree of unpredictability.

Change and novelty as stress inducing agents can be studied through research designs providing for careful measurement or systematic variation in the physical and human environments, i.e., the degree of carryover of familiar objects from the old to the new environment, the degree of similarity between the old and new caretakers in physical and psychological characteristics, variations among the old and new mothers in the modalities in which stimulation is given. The impact of change in the physical environment might be evaluated by holding constant the human environment while systematically varying the physical environment, and conversely, the human environment might be varied, with the physical environment constant. The amount of change necessary to produce a discriminable difference to the child may vary with developmental factors. The significance of a change in the human environment will almost certainly depend on whether a meaningful relationship has developed with the mother-figure. If separation occurs after this point, the stress of change is reinforced by the loss of a significant person.

In the research on multiple mothering the one consistent characteristic of the varied contexts of multiple mothering is environmental unpredictability associated with changing agents of gratification. Unpredictability may be based on differences in technique among the different mother-figures, on variations in speed of response to the child's expression of needs, on inconsistency in the kinds of behavior which are rewarded, punished, or ignored. Unlike separation conditions in which new predictable patterns may soon be established, in multiple mothering unpredictability remains the most characteristic aspect of the environment.

There is not strong research evidence nor very firm theoretical grounds to support the assumption that the presence of several concurrent mother-figures in early life results in a diffusion of the mother-image and later inability to establish meaningful relationships. The variable conditions of reinforcement which characterize some multiple mothering situations provide a special kind of learning situation which may lead to the development of atypical patterns of relationships, but not necessarily shallow ones. It is likely that the presence of several mother-figures will vary in significance at different developmental periods. The lack of a consistent role model is probably more serious during the early preschool period than in early infancy. In further research, attempts should be made to vary systematically the degree of stress associated with environmental unpredictability, while controlling other variables such as degree of role differentiation among the multiple mothers.

Although deprivation and trauma can be treated as independent concepts, there are conditions under which deprivation can be considered a traumatic stimulus. It is recognized that trauma may result from excessive stimulation, but the conditions under which inadequate stimulation may be traumatic are more obscure. Recent research indicates that extreme sensory deprivation may be stressful for adults (Wexler, Mendelson, Leiderman, & Solomon, 1959). We might assume that deprivation becomes a traumatic stimulus after the appropriate motivational conditions have developed. Thus Hebb (1955) suggests:

> The observed results seem to mean, not that the stimulus of another attentive organism (the mother) is necessary from the first, but that it may become necessary only as psychological dependence on the mother develops (p. 828).

Research Implications

Analysis of the research on institutionalization, separation, and multiple mothering highlights some theoretically significant questions and points to some specific variables which can be experimentally manipulated or controlled through the opportunistic utilization of natural situations.

Duration of deprivation or stress. In much of the research, the subjects have experienced a cumulative series of deprivations or stressful experiences, beginning in infancy and continuing through childhood. Few studies give specific data on the length of time the child has been exposed to these conditions. Goldfarb (1945b, 1947, 1955), Bender (1945, 1947), and Bowlby (1944) conclude from retrospective studies that the longer the period of institutional care, the more severe the ultimate damage. These conclusions are based largely on individual case findings. Those cases which did not show the same irreversible patterns as the rest of the population had been in institutions for a shorter period of time. Spitz and Wolf (1946) suggest that there may be a critical time interval after which the effects of maternal separation are irreversible.

If the infant is reunited with his mother within 3 months, the process of physical, social, and intellectual deterioration may be arrested, but if the mother-child relationship is not restored within 5 months, irreparable damage occurs. There are no comparable data on children beyond infancy. One might hypothesize that the critical time interval might be longer with older children.

Research on older children attest to the damaging effects of repeated separations (Bowlby, 1944; Lewis, 1954). On the whole, no distinction has been made among several different separation experiences: a single instance of separation with reunion, a single separation without reunion, repeated small doses of separation with consistent reunion with the same mother, and cumulative separations with repeated changes in mothers. It can be assumed that each of these experiences provides different learning conditions for the development of meaningful relationships. The most extreme outcome, the "affectionless character," may be the result of the most extreme conditions, i.e., repeated traumatic separations.

Time or developmental stage at which deprivation or stress occurs. Psychoanalytic theories regarding the significance of early experience for later development have often been interpreted as postulating that the younger the organism, the more severe and fixed the effects of an environmental impact. Only limited data are available on human subjects. Ribble (1943) tends to interpret her data on maternal rejection as supporting this point of view. Bender's and Goldfarb's (1947) retrospective studies suggest that the younger the child, the more damaging the effects of deprivation and stress. Some animal research supports this hypothesis; other studies do not (Beach & Jaynes, 1954; King, 1958).

The findings on institutionalized infants that intellectual retardation is not apparent before 3 months of age and that personality disturbances are not evident before 5 or 6 months suggest that this type of deprivation has no significant impact in the early weeks of infancy. (Because of the known unreliability of infant tests, and the lack of sensitive measures of personality and intellectual functions in early infancy, some degree of caution is necessary in interpreting these findings.)

A more refined hypothesis regarding the significance of the timing of experiences is the critical phase hypothesis which holds that there are points in the developmental cycle during which the organism may be particularly sensitive to certain kinds of events or most vulnerable to specific types of deprivation or stress. Several animal studies (Moltz, 1960; Scott, Fredericson, & Fuller, 1951; Tinbergen, 1954) support the general outlines of the critical phase hypothesis. From the assorted data on the intellectual functioning of institutionalized children a testable hypothesis emerges regarding a critical period for institutional deprivation: vulnerability to intellectual damage is greatest during the 3–12 month period. Beres and Obers (1950) suggest that institutional deprivation will differ in its impact at different developmental periods. The data on which this conclusion is based are limited. Of their four cases showing mental retardation, all were admitted to the institution under 6 months of age; the four cases developing schizophrenia entered the institution at a later age (specific age not reported).

Although the general consensus in the literature is that maternal separation which occurs before the child is 5 years of age is likely to be most damaging, the findings are not sufficiently clear to pinpoint any one age as being most vulnerable. Bowlby (1944) notes among the affectionless thieves:

> in practically all these cases, the separation which appears to have been pathogenic occurred after the age of six months, and in a majority after twelve months. This suggests that there is a lower age limit, before which separations, whilst perhaps having undesirable effects, do not produce the particular results we are concerned with here—the affectionless and delinquent character (p. 41). (Quoted by permission of the *International Journal of Psycho-Analysis.*)

On the basis of our knowledge of the developmental characteristics of children, one might postulate differing vulnerabilities at different periods of development. The developmental level of the child is likely to influence the significance of deprivation or the meaning of a separation experience for him. With regard to separation, the period during which the child is in the process of consolidating a relationship with his mother may be an especially vulnerable one. Also significant may be the developmental stage with regard to memory functions. After the point in development at which the child can sustain an image of the mother in her absence and can anticipate her return, the meaning of a brief separation may be less severe than at an earlier developmental period. The degree of autonomy the child has achieved may also affect the extent of trauma experienced. The loss of the mother may represent a greater threat to the completely dependent infant than to the young child who has achieved some locomotion and some manipulatory control over his environment. The advent of language which symbolizes even a greater degree of environmental mastery may mitigate further the severity of trauma.

Similarly, the effects of institutional deprivation may be more severe for the young infant who is completely dependent on outside sources of stimulation than for the older child who is capable of seeking out stimulation. There may also be age linked effects of different types of deprivation. Some animal studies suggest that a minimal level of stimulation may be necessary to produce the biochemical changes necessary for the development of the underlying structures. Deprivation in certain sensory modalities may be more significant at one age than at another. For example, deprivation of tactile stimulation may be most significant during the first weeks of infancy, whereas auditory or visual deprivation may become more significant later. Social deprivation may be most damaging during the earliest period of the development of social responsiveness.

Constitutional factors. Although the role of constitutional factors in influencing the long-term effects of early trauma has been increasingly stressed, the meager data in support of the significance of constitutional factors have been indirect. Several retrospective studies have found similar deprivation experiences in the history of individuals who in later life made satisfactory life adjustments as in those who made poor adjustments. The different outcomes are accounted for in terms of constitutional factors. In considering the role of constitutional factors a distinction might be made between organismic differences in general vulnerability to deprivation or stress and vulnerabilities in specific sensory modalities.

Data from a number of studies attest to individual differences in sensitivities in specific modalities. With regard to research design, it may be important, too, to distinguish between organismic differences which are constitutionally determined and differences in vulnerability which vary with developmental stage. While organismic sensitivities cannot be manipulated experimentally, it may be possible to study constitutional factors by developing research designs in which subjects with known differences in sensitivities are subjected to the same experimental conditions.

The Long-Term Effects: The Issue of Reversibility

It does not seem fruitful to state the question of reversibility in terms of an either-or hypothesis, i.e., whether or not early experiences produce irreversible effects. Rather the question might be: what are the conditions under which an earlier traumatic or depriving experience is likely to produce irreversible effects? The concept of irreversibility implies that an adverse experience results in permanent structural changes in the nervous system such that at some later developmental period a given response sequence is either facilitated or inhibited. A further implication is that subsequent experience plays no role in changing response potentialities or in developing responses which are incompatible with earlier established behavior patterns. Several studies suggest that permanent damage to the central nervous system may result from early sensory deprivation. Increasingly the research points to the resiliency of the organism. Beres and Obers' is one of the few investigations from the psychoanalytic orientation which makes a strong case for the modifiability of the effects of earlier infantile experience. They cite in support a conclusion by Hartmann, Kris, and Lowenstein (1946) that

> the basic structure of the personality and the basic functional interrelationship of the systems of the ego and superego are fixed to some extent by the age of six, but after this age, the child does not stop growing and developing, and growth and development modify existing structure (p. 34). (Quoted by permission of International Universities Press, Inc.)

Many factors in complex interaction undoubtedly determine the extent to which recovery is possible from early intellectual or personality damage. More pointed research is needed to identify the specific conditions under which irreversible damage to the central nervous system occurs. Also needed are specific research designs on reversibility, designs aimed at reversing intellectual or personality damage.

Toward a Concept of Maternal Deprivation

In focusing on the isolation of simple variables for formulating testable hypotheses on the relationship between early environmental conditions and later development, we have avoided complex concepts centering around the emotional interchange between mother and infant, concepts which have been focal in psychodynamic theories. The mother as a social stimulus provides sensory stimulation to the infant through tactile, visual, and auditory media, i.e., through handling, cuddling, talking and playing with the child, as well as by simply being visually present. The mother also acts as a mediator of environmental stimuli, bringing the infant in contact with the environment and buffering or heightening the intensity of stimuli. The meaning of these mothering activities to the child and the impact of the mother's absence varies with the child's perceptual, cognitive, and motor capacities at different developmental levels. On the simplest level, if the mother is not present, the infant may be deprived of tactile, auditory, and visual stimuli from a social source, as well as of the environmental stimuli which the mother ordinarily makes available to him. At this point, the mother's absence may be experienced by the young infant only as a deprivation of distinctive stimuli offered by a social being. The impact on the infant may be more severe if the mother's absence is accompanied by deviations in need-gratification sequences, such as, failure to have needs anticipated or long delay before gratification is provided, by marked inconsistencies in patterns of gratification, or inadequate gratification. The significance of these kinds of frustration experiences will be modified by the length of time during which they operate, the developmental level of the child, e.g., the degree of autonomy he has achieved.

The usefulness of this reduction of maternal deprivation has been demonstrated in ordering the reported research findings and in suggesting more refined hypotheses for further research. It is likely, however, that not all aspects of the mother-child relationship can be meaningfully reduced to such simple variables. We can only speculate on the process through which the mother comes to acquire special meaning to the child. We assume that the mother-image gradually evolves as a distinctive perceptual entity out of a welter of tactile, visual, auditory, and kinesthetic cues. (There has been some speculation, without definitive data, that in early infancy before these sensory cues are organized into a percept of an object existing outside of himself, the infant may still "recognize" the mother as an assortment of familiar stimuli.) In time through repeated contact these cues become "familiar" or distinctive to the infant, and finally there is a fixation of positive feelings on this perceptual complex. After the point of fixation of positive feelings on the mother, new elements enter into the child's reactions to a loss or a change in mothers. At this point, sensory deprivation and environmental change may be secondary, the loss of a significant person becomes of primary significance. This experience cannot occur until the infant reaches a developmental point at which he is able to conceptualize the existence of an "object" outside of himself. As a matter of conceptual clarity, it might be desirable to limit the concept of maternal deprivation to the conditions associated with the loss of a specific, cathected person, a person who has acquired distinctive significance for the child, one on whom positive feelings have been fixated.

Conclusions

The wide range of circumstances included under the concept of maternal deprivation stand out when the research is carefully scrutinized. Included are studies of children who have been separated from their parents and placed in institutional settings, other studies deal with children who have been grossly maltreated or rejected by their families, others are concerned with children temporarily separated from their parents because of illness, and in others the maternal functions are assumed by several different persons. These experiences have occurred at different developmental stages in the children's life histories, and there has been considerable variation in the length of exposure to these conditions, and in the circumstances preceding and following the deviating conditions.

It is apparent that the data on maternal deprivation are based on research of varying degrees of methodological rigor. Most of the data consist of descriptive clinical findings arrived at fortuitously rather than through planned research, and frequently the findings are based on retrospective analyses which have been narrowly directed toward verification of clinical hunches.

The areas of knowledge and the areas of uncertainty become more sharply delimited when we break down the complex concept of maternal deprivation into some discrete variables. For instance, in the studies on institutional care in which sensory deprivation emerges as a major variable, we can conclude that severe sensory deprivation before one year of age, if it continues for a sufficiently long period of time, is likely to be associated with severe intellectual damage. Direct observation of children undergoing the experience of maternal separation shows a variety of immediate disturbances in behavior, permitting the simple conclusion that this is a stressful experience for children. There is no clear evidence that multiple mothering, without associated deprivation or stress, results in personality damage.

With regard to the long-term effects of early deprivation or stress associated with institutionalization or maternal separation, no simple conclusions can be drawn. In the retrospective studies, significant interacting variables are usually unknown. Longitudinal studies currently underway may offer data on the reinforcing or attenuating influence of later experiences. We might hope for more pointed longitudinal studies on questions of reversibility, such as, studies of human or animal subjects who have been subjected to experimental deprivation or trauma, or longitudinal studies of special populations chosen because of some known deviation from a cultural norm of mothering, e.g., infants who have experienced separation for adoption (Yarrow, 1955, 1956) and infants in multiple mothering situations (Pease & Gardner, 1958).

The analysis of the literature points up the need for more definitive research on the role of many "nonmaternal" variables, variables relating to the characteristics of environmental stimulation and variables dealing with organismic sensitivities. After clarification of the influence of such variables, then perhaps systematic research can come to grips with some of the more elusive aspects of the emotional interchange in the intimate dyadic relationship of mother and infant.

REFERENCES[2]

AINSWORTH, MARY D., & BOSTON, MARY. Psychodiagnostic assessment of a child after prolonged separation in early childhood. *Brit. J. med. Psychol.*, 1952, **25**, 169–205.

AINSWORTH, MARY D., & BOWLBY, J. Research strategy in the study of mother-child separation. *Courr. Cent. Int. l'Enfance*, 1954, **4**, 1–47.

BAKWIN, H. Emotional deprivation in infants. *J. Pediat.*, 1949, **35**, 512–521.

BEACH, F. A., & JAYNES, J. Effects of early experience upon the behavior of animals. *Psychol. Bull.*, 1954, **51**, 239–263.

[2] Due to space limitations, many relevant references have not been cited. An extensive bibliography of earlier studies can be found in Bowlby (1951).

BENDER, LAURETTA. Infants reared in institutions: Permanently handicapped. *Bull. Child Welf. League Amer.*, 1945, 24, 1–4.

BENDER, LAURETTA. There's no substitute for family life. *Child Stud.*, 1946, 23, 74–76, 96.

BENDER, LAURETTA, Psychopathic behavior disorders in children. In R. M. Linder (Ed.), *Handbook of correctional psychology.* New York: New York Philosophical Library, 1947. Pp. 360–377.

BENDER, LAURETTA, & YARNELL, H. An observation nursery: A study of 250 children in the psychiatric division of Bellevue Hospital. *Amer. J. Psychiat.*, 1941, **97**, 1158–1174.

BERES, D., & OBERS, S. J. The effects of extreme deprivation in infancy on psychic structure in adolescence. *Psychoanal. Stud. Child*, 1950, **5**, 121–140.

BERG, M., & COHEN, B. B. Early separation from mother in schizophrenia. *J. nerv. ment. Dis.*, 1959, 128, 365–369.

BODMAN, F., et al. The social adaptation of institution children. *Lancet*, 1950, 258, 173–176.

BOWLBY, J. Forty-four juvenile thieves. *Int. J. Psycho-Anal.*, 1944, **25**, 1–57.

BOWLBY, J. Maternal care and mental health. *WHO Monogr.*, 1951, No. 2.

BOWLBY, J. Some pathological processes engendered by early mother-child separation. In M. J. Senn (Ed.), *Infancy and childhood.* New York: Josiah Macy, Jr. Foundation, 1953, Pp. 38–87. (a)

BOWLBY, J. Some pathological processes set in train by early mother-child separation. *J. ment. Sci.*, 1953, 99, 265–272. (b)

BOWLBY, J. An ethological approach to research in child development. *Brit. J. med. Psychol.*, 1957, **30**, 230–240.

BOWLBY, J. The nature of the child's tie to the mother. *Int. J. Psycho-Anal.*, 1958, **39**, 1–24.

BOWLBY, J., AINSWORTH, MARY, BOSTON, MARY, & ROSENBLUTH, DINA. The effects of mother-child separation: A follow-up study. *Brit. J. med. Psychol.*, 1956, 29, 211–247.

BRODBECK, A. J., & IRWIN, O. C. The speech behavior of infants without families. *Child Develpm.*, 1946, 17, 145–156.

BROWN, F. Neuroticism of institution vs. non-institution children. *J. appl. Psychol.*, 1937, 21, 379–383.

COLEMAN, RUTH W., & PROVENCE, SALLY. Environmental retardation (hospitalism) in infants living in families. *Pediatrics*, 1957, 19, 285–292.

DAY, ELLA J. The development of language in twins: I. A comparison of twins and single children. *Child Develpm.*, 1932, 3, 179–199.

DENNIS, W. Infant development under conditions of restricted practice and of minimum social stimulation. *Genet. psychol. Monogr.*, 1941, 23, 143–190.

DENNIS, W., & DENNIS, MARSENA G. Development under controlled environmental conditions. In W. Dennis (Ed.), *Readings in child psychology.* New York: Prentice-Hall, 1951. Pp. 104–131.

DENNIS, W., & NAJARIAN, P. Infant development under environmental handicap. *Psychol. Monogr.*, 1957, 71(7, Whole No. 436).

DUBOIS, CORA. *The people of Alor.* Minneapolis: Univer. Minnesota Press, 1944.

DUPAN, R. M., & ROTH, S. The psychologic development of a group of children brought up in a hospital type residential nursery. *J. Pediat.*, 1955, **47**, 124–129.

EDELSTON, H. Separation anxiety in young children: A study of hospital cases. *Genet. psychol. Monogr.*, 1943, 28, 3–95.

EGGAN, D. The general problem of Hopi adjustment. *Amer. Anthropologist*, 1945, **47**, 516–539.

FISCHER, LISELOTTE. Hospitalism in six month old infants. *Amer. J. Orthopsychiat.*, 1952, 22, 522–533.

FISCHER, LISELOTTE. Psychological appraisal of the unattached preschool child. *Amer. J. Orthopsychiat.*, 1953, 23, 803–814.

FLINT, BETTY. Babies who live in institutions. *Bull. Inst. Child Stud., Toronto*, 1957, 19, 1–5.

FREUD, ANNA, & BURLINGHAM, DOROTHY T. *Infants without families.* New York: International Univer. Press, 1944.

GARDNER, D. B., PEASE, DAMARIS, & HAWKES, G. R. Responses of two-year-old adopted children to controlled stress situations. Paper read at Society for Research in Child Development, Washington, D. C., March 1959.

GARDNER, D. B., & SWIGER, M. K. Developmental status of two groups of infants released for adoption. *Child Develpm.*, 1958, 29, 521–530.

GESELL, A., & AMATRUDA, CATHERINE. *Developmental diagnosis.* New York: Hoeber, 1941.

GEWIRTZ, J. L. Social deprivation and dependency: A learning analysis. Paper read in symposium on Dependency in personality development, American Psychological Association, New York, August 1957.

GLASER, K., & EISENBERG, L. Maternal deprivation. *Pediatrics*, 1956, 18, 626–642.

GOLAN, S. Behavior research in collective settlements in Israel: Collective education in the kibbutz. *Amer. J. Orthopsychiat.*, 1958, **28**, 549–556.

Goldfarb, W. Effects of early institutional care on adolescent personality (graphic Rorschach data). *Child Developm.*, 1943, **14**, 213–223. (a)

Goldfarb, W. Infant rearing and problem behavior. *Amer. J. Orthopsychiat.*, 1943, **13**, 249–265. (b)

Goldfarb, W. Effects of early institutional care on adolescent personality: Rorschach data. *Amer. J. Orthopsychiat.*, 1944, **14**, 441–447. (a)

Goldfarb, W. Infant rearing as a factor in foster home replacement. *Amer. J. Orthopsychiat.*, 1944, **14**, 162–173. (b)

Goldfarb, W. Effects of psychological deprivation in infancy and subsequent stimulation. *Amer. J. Psychiat.*, 1945, **102**, 18–33. (a)

Goldfarb, W. Psychological privation in infancy and subsequent adjustment. *Amer. J. Orthopsychiat.*, 1945, **15**, 247–255. (b)

Goldfarb, W. Variations in adolescent adjustment of institutionally reared children. *Amer. J. Orthopsychiat.*, 1947, **17**, 449–457.

Goldfarb, W. Rorschach test differences between family-reared, institution-reared, and schizophrenic children. *Amer. J. Orthopsychiat.*, 1949, **19**, 625–633.

Goldfarb, W. Emotional and intellectual consequences of psychologic deprivation in infancy: A re-evaluation. In P. H. Hoch & J. Zubin (Eds.), *Psychopathology of childhood.* New York: Grune & Stratton, 1955. Pp. 105–119.

Haggerty, A. D. The effects of long-term hospitalization upon the language development of children. *J. genet. Psychol.*, 1959, **94**, 205–209.

Harlow, H. The nature of love. *Amer. Psychologist*, 1958, **15**, 673–685.

Hartmann, H., Kris, E., Lowenstein, R. M. Comments on the formation of psychic structure. *Psychoanal. Stud. Child*, 1946, **2**, 11–38.

Hebb, D. O. *The organization of behavior: A neuropsychological theory.* New York: Wiley, 1949.

Hebb, D. O. The mammal and his environment. *Amer. J. Psychiat.*, 1955, **3**, 826–831.

Heinicke, C. Some effects of separating two-year-old children from their parents: A comparative study. *Hum. Relat.*, 1956, **9**, 105–176.

Irvine, Elizabeth. Observations on aims and methods of child rearing in communal settlements in Israel. *Hum. Relat.*, 1952, **5**, 247–275.

Kardiner, A. Social stress and deprivation. In I. Galdston (Ed.), *Beyond the germ theory.* New York: New York Health Education Council, 1954. Pp. 147–170.

King, J. A. Parameters relevant to determining the effect of early experience upon the adult behavior of animals. *Psychol. Bull.*, 1958, **55**, 46–58.

Klackenburg, G. Studies in maternal deprivation in infant homes. *Acta paediat., Stockh.*, 1956, **45**, 1–12.

Levy, D. Primary affect hunger. *Amer. J. Psychiat.*, 1937, **94**, 643–652.

Levy, Ruth. Institutional vs. boarding-home care. *J. Pers.*, 1947, **15**, 233–241.

Lewis, Hilda. *Deprived children.* Toronto: Oxford Univer. Press, 1954.

Lowrey, L. G. Personality distortion and early institutional care. *Amer. J. Orthopsychiat.*, 1940, **10**, 576–585.

Mead, Margaret. *Sex and temperament in three primitive societies.* New York: Mentor, 1935.

Moltz, H. Imprinting: Empirical basis and theoretical significance. *Psychol. Bull.*, 1960; **57**, 291–314.

Pease, Damaris, & Gardner, D. B. Research on the effects of non-continuous mothering. *Child Develpm.*, 1958, **29**, 141–148.

Pinneau, S. R. A critique on the articles by Margaret Ribble. *Child Develpm.*, 1950, **21**, 203–228.

Pinneau, S. R. The infantile disorders of hospitalism and anaclitic depression. *Psychol. Bull.*, 1955, **52**, 429–462.

Rabin, A. I. Personality maturity of kibbutz (Israeli collective settlement) and non-kibbutz children as reflected in Rorschach findings. *J. proj. Tech.*, 1957, **31**, 148–153.

Rabin, A. I. Behavior research in collective settlements in Israel: Infants and children under conditions of "intermittent" mothering in the kibbutz. *Amer. J. Orthopsychiat.*, 1958, **28**, 577–586. (a)

Rabin, A. I. Some psychosexual differences between kibbutz and non-kibbutz Israeli boys. *J. proj. Tech.*, 1958, **22**, 328–332. (b)

Rapaport, D. Behavior research in collective settlements in Israel: The study of Kibbutz education and its bearing on the theory of development. *Amer. J. Orthopsychiat.*, 1958, **28**, 587–597.

Rheingold, Harriet L. Mental and social development of infants in relation to the number of other infants in the boarding home. *Amer. J. Orthopsychiat.*, 1943, **13**, 41–44.

Rheingold, Harriet L. The modification of social responsiveness in institutional babies.

Monogr. Soc. Res. Child Developm., 1956, 21, No. 63.

RHEINGOLD, HARRIET L. The measurement of maternal care. *Child Develpm.*, 1960, **31**, 565–573.

RHEINGOLD, HARRIET L., & BAYLEY, N. The later effects of an experimental modification of mothering. *Child Developm.*, 1959, **30**, 363–372.

RHEINGOLD, HARRIET L., GEWIRTZ, J., & ROSS, HELEN. Socialconditioningof vocalizations in the infant. *J. comp. physiol. Psychol.*, 1959, 52, 58–73.

RIBBLE, MARGARET. *Rights of infants.* New York: Columbia Univer. Press, 1943.

ROBERTSON, J., & BOWLBY, J. Responses of young children to separation from their mothers: II. Observation of sequences of response of children aged 18–24 months during course of separation. *Courr. Cent. Int. l'Enfance*, 1952, 2, 131–139.

ROSCOE, J. Baganda: An account of their native customs and beliefs. In I. T. Sanders (Ed.), *Societies around the world.* New York: Dryden, 1953. Pp. 412–420.

ROUNDINESCO, JENNY, DAVID, MIRIAM, & NICOLAS, J. Responses of young children to separation from their mothers: I. Observation of children ages 12 to 17 months recently separated from their families and living in an institution. *Courr. Cent. Int. l'Enfance*, 1952, 2, 66–78.

SCHAFFER, H. R. Objective observations of personality development in early infancy. *Brit. J. med. Psychol.*, 1958, **31**, 174–183.

SCOTT, J. P., FREDERICSON, E., & FULLER, J. L. Experimental exploration of the critical period hypothesis. *Personality*, 1951, **1**, 162–183.

SKEELS, H. M. A study of the effects of differential stimulation on mentally retarded children: Follow-up report. *Amer. J. ment. Defic.*, 1942, **66**, 340–350.

SKEELS, H. M., & DYE, H. A study of the effects of differential stimulation on mentally retarded children. *Proc. Amer. Ass. Ment. Defic.*, 1939, 44, 114–136.

SKEELS, H. M., UPDEGRAFF, RUTH, WELLMAN, BETH L., & WILLIAMS, H. M. A study of environmental stimulation: An orphanage preschool project. *U. Ia. Stud. child Welf.*, 1938, **15**, 7–191.

SPITZ, R. A. Hospitalism: An inquiry into the genesis of psychiatric conditions in early childhood. *Psychoanal. Stud. Child.*, 1945, **1**, 53–74; 1946, **2**, 113–117.

SPITZ, R. A. The psychogenic diseases in infancy: An attempt at their etiologic classification. *Psychoanal. Stud. Child*, 1951, **6**, 255–275.

SPITZ, R. A. Infantile depression and the general adaptation syndrome. In P. H. Hoch & J. Zubin (Eds.), *Depression.* New York: Grune & Stratton, 1954. Pp. 93–108. (a)

SPITZ, R. A. Unhappy and fatal outcomes of emotional deprivation and stress in infancy. In I. Galdston (Ed.), *Beyond the germ theory.* New York: New York Health Education Council, 1954. Pp. 120–131. (b)

SPITZ, R. A. Reply to Pinneau. *Psychol. Bull.*, 1955, **52**, 453–459.

SPITZ, R. A., & WOLF, KATHERINE. Anaclitic depression. *Psychoanal. Stud. Child*, 1946, **2**, 313–342.

SPITZ, R. A., & WOLF, KATHERINE. Autoerotism. *Psychoanal. Stud. Child*, 1949, **3–4**, 85–120.

STERN, W. *General psychology from the personalistic standpoint.* New York: Macmillan, 1938.

STONE, L. J. A critique of studies of infant isolation. *Child Develpm.*, 1954, **25**, 9–20.

TINBERGEN, N. Psychology and ethology as supplementary parts of a science of behavior. In B. Schaffner (Ed.), *Group processes.* New York: Josiah Macy, Jr. Foundation, 1954.

VAN ALSTYNE, D. The environment of three-year-old children: Factors related to intelligence and vocabulary tests., *Teach. Coll. Contr. Educ.*, 1929, No. 366.

WELLMAN, BETH, & PEGRAM, E. L. Binet IQ changes of orphanage preschool children: A re-analysis. *J. genet. Psychol.*, 1944, **65**, 239–263.

WEXLER, D., MENDELSON, J., LIEDERMAN, P. H., & SOLOMON, P. Sensory deprivation. *AMA Arch. Neurol. Psychiat.*, 1958, **79**, 225–233.

YARROW, L. J. Research on maternal deprivation. Paper read at symposium on Maternal deprivation, American Association for the Advancement of Science, Section I, Atlanta, Georgia, December 1955.

YARROW, L. J. The development of object relationships during infancy, and the effects of a disruption of early mother-child relationships. *Amer. Psychologist*, 1956, **11**, 423. (Abstract)

(Received September 12, 1960)

Psychological Bulletin
1961, Vol. 58, No. 6, 491–514

THE CURRENT STATUS OF THE SIZE-DISTANCE HYPOTHESES[1]

WILLIAM EPSTEIN, JOHN PARK, AND ALBERT CASEY
University of Kansas

In the history of the psychology of perception few matters have been of more continuous interest than the relationship between perceived size and perceived distance. It is our objective to examine the current status of this question by reviewing the recent literature. With some exceptions our review will be confined to investigations which have been reported since 1952. Several surveys of the literature prior to 1952 are available, and for this reason we will have relatively little to say about these earlier investigations (reviews can be found in Boring, 1942, Ch. 8; Vernon, 1954, Ch. 5; Woodworth & Schlosberg, 1954, Ch. 16).[2]

Most studies of this question have converged upon a single proposition which aptly has been called the Size-Distance Invariance Hypothesis. The invariance hypothesis is often stated in the following terms: "A retinal projection or visual angle of given size determines a unique ratio of apparent size to apparent distance" (Kilpatrick & Ittelson, 1953, p. 224). This proposition has been applied repeatedly in explanations of perceived size and distance in general, and in accounts of size constancy in particular.

Two variations of this fundamental proposition also have been asserted frequently. The first may be called the Known Size-Apparent Distance Hypothesis, and it can be derived directly from the more general proposition stated above. It may be expressed as follows: an object of known physical size uniquely determines the relation of the subtended visual angle to apparent distance. This hypothesis is the basis for many explanations of size as a cue for apparent distance.

The second variation is often called Emmert's Law, and in this form has been employed in investigations of the size of the afterimage and its relationship to the distance of the projection surface. Woodworth and Schlosberg have stated the relationship in this way: "the judged size of the image is proportional to the distance" (1954, p. 486). A more general statement can be formulated also: the apparent size of an object will be proportional to distance when retinal size is constant. In this form the close relationship between this proposition and the broader Size-Distance Invariance Hypothesis is obvious. We have given the proposition independent status because it has been applied mainly to questions concerned with the perceived size of the afterimage.

For clarity of exposition we have elected to review each of these propositions separately. However, the

[1] Work on this paper was supported by a grant (M-4153) to the first author from the United States Public Health Service, a grant from the General Research Fund of the University of Kansas, and a predoctoral research fellowship awarded to the second author by the United States Public Health Service.

[2] Several reviews which have appeared more recently have not added very much to the earlier treatments (see Bartley, 1958, pp. 179–187; Dember, 1960, pp. 169–192). The same can be said about the presentations contained in the recently published opthalmological textbooks. Two illustrative discussions can be found in Bedrossian (1958, pp. 109–115) and Adler (1959, pp. 762–780).

reader will discover that on several occasions we have violated these self-imposed boundaries. In the closing section of this paper we shall present some conclusions about the size-distance relationship in general.[3]

THE SIZE-DISTANCE INVARIANCE HYPOTHESIS

This hypothesis proposes an invariant relationship between perceived size and distance such that the apparent size of an object is uniquely determined by an interaction of visual angle and apparent distance.

Support for the invariance hypothesis comes from studies which show that the size of an unfamiliar object can be judged accurately only if cues to the distance of the object are available. The prototypal experiment was performed by Holway and Boring (1941), who obtained size matches under four sets of conditions which represented a successive elimination of distance cues. Size matches approximated constancy under conditions of binocular viewing and gradually approached the law of visual angle as distance cues were eliminated. However, perfect visual angle matches were not obtained even under the condition of greatest reduction. This was attributed to a "light haze" visible within the reduction tunnel due to light reflections in the corridor. When this cue was eliminated, perfect visual angle matches were obtained (Lichten & Lurie, 1950). These findings have been confirmed in more recent investigations which utilized a variety of stimulus objects and a variety of techniques for eliminating distance cues (e.g., Chalmers, 1952, 1953; Hastorf & Way, 1952; Renshaw, 1953; Zeigler & Leibowitz, 1957).

The results referred to above are usually interpreted as a straightforward demonstration of the dependence of perceived size on perceived distance. However, we wish to point out that the introduction of the visual angle matches as evidence for the size-distance hypothesis involves at least one of the following two assumptions: (*a*) under conditions of complete reduction apparent distance tends toward zero, (*b*) under conditions of complete reduction apparent distance assumes some value other than zero which is the same for both the standard and the variable stimulus.

The first assumption is untenable in its original form since the value "zero" distance is meaningless in the experimental contexts described earlier. Perhaps, then, "zero distance" might be interpreted to mean indeterminate distance, i.e., distance which is not regulated by specifiable cues. Still, as Woodworth and Schlosberg note, "we just do not perceive free-floating objects at unspecified distances" (1954, p. 481). Instead, the object will be localized at some specific distance. According to the invariance hypothesis, the apparent distance for any given observer (*O*), whatever it is, should interact with the visual angle to determine apparent size. However, since the reduced situation is ambiguous it is likely that apparent

[3] Various areas of relevant research have been omitted from this paper. Investigations dealing with the relationship between exposure time and perceived size (e.g., Allen, 1953; Comalli, 1951; Gulick & Stake, 1957; Howarth, 1951; Leibowitz, Chinetti, & Sidowsky, 1956) and the effects of relative visual direction on perceived size and distance (e.g., Gogel, 1954, 1956a, 1956b) have not been reviewed. We have also excluded reference to the developmental studies of size and distance. These investigations have been reviewed recently by Wohlwill (1960).

distance will vary for different *Os*. Under these conditions, the invariance hypothesis would predict corresponding variations in the size matches. This prediction, of course is quite different from the consistent visual angle matches obtained by Holway and Boring, etc. For these reasons the first assumption stated in terms of "zero" distance or "indeterminate" distance is not very convincing to us.

The assumption of equidistance seems more plausible. Carlson (1960a) and Wallach and McKenna (1960), addressing themselves to different aspects of the size-distance problem, have advanced the second assumption. Thus, Wallach and McKenna write that "the equation of image-sizes results from an implicit assumption of equal distance of the standard and the comparison object" (1960, p. 460). Carlson (1960a, p. 14) cites Gogel's (1956b) experiments as evidence for a tendency to see objects as equidistant under the conditions of the reduction experiment.

It is plain that a bias toward equidistance would explain the obtained visual angle matches. Unfortunately, there is little empirical basis for the contention that this tendency actually was operative. The experimental evidence for the equidistance tendency (Judd, 1898; Gogel, 1956b) was obtained when all of the objects in question were viewed simultaneously. In the classic Holway-Boring investigation the standard and comparison were viewed successively. Secondly, all of Gogel's experiments dealt with instances in which a monocularly viewed object was localized at the same distance as a *binocularly* viewed object. Gogel presented no evidence that the same equidistance tendency is present when all objects were viewed monocularly. However, the Holway-Boring results were obtained when both standard and comparison were viewed monocularly. Finally, it should be noted "that the strength of the tendency for objects to appear equidistant decreases as the lateral line-of-sight separation of the objects is increased" (Gogel, 1956b, p. 16). This fact makes it highly unlikely that the equidistance tendency was effective in the Holway-Boring type of experiment.

This analysis leads us to conclude that the applicability of the visual angle data as evidence for the invariance hypothesis involves assumptions whose validity has never been demonstrated. What is needed is a systematic experimental investigation of apparent distance under varying degrees of reduction including complete elimination of distance cues. In the absence of such information the consonance of visual angle matches with the invariance hypothesis is at best conjectural.

The frequent appeals to the invariance hypothesis in explanations of perceived size have endowed this proposition with almost axiomatic status. Nonetheless, evidence has been accumulating which casts doubt on the generality of this hypothesis. In what follows we shall describe a series of investigations whose outcomes have not been consonant with the invariance hypothesis.

Overestimation in Size Judgments

A frequently confirmed finding is size overestimation which increases with distance. As the physical distance of the object is increased, the physical size of the object is progressively overestimated. While overestimation is certainly surprising, it need not necessarily be inconsistent with the invariance hypothesis. If it should also turn out that apparent distance increases more rapidly than physical distance, then the results demonstrating increasing overestimation of size could be reconciled easily.

Let us first consider those studies which report instances of overestimation of size which increases with distance. Unless otherwise noted, the results to be described below were obtained with binocular vision and an objective matching attitude, i.e., *O* was instructed to match the standard and comparison so that they would have the same physical size. Holway and Boring (1941) found that when *O* was allowed normal binocular vision, the apparent size of a disk of light increased more rapidly with increasing physical distance than did physical size. This finding was explained as a "space error" resulting from the fact that the variable stimulus was always to the left of the standard. More recent experiments rule out this explanation. In an outdoor setting, Gibson (1947, 1950) had *O* match the size of a distant stake with the size of one of a set of nearer stakes, which stood both to the right and to the left of the more distant stake. Overestimation of the size of the distant stake increased as its distance increased from approximately 80 feet to 675 feet. The increase of estimated size with distance was greatest between 80 and 320 feet.

More recent experiments confirm Gibson's findings. Gilinsky (1955a) investigated size perception of objects presented out-of-doors at distances ranging from 100 to 4,000 feet. Size matches made under an "objective" set were greater than the physical size of the standards and increased with increasing distance of the standard. The acceleration of estimated size with distance was greatest between 100 and 400 feet. Using somewhat shorter distances and three-dimensional stimulus objects in an outdoor setting, Smith (1953) also demonstrated that apparent size increases with distance. Under Distance Condition N, the comparison was placed at 2 feet and the standard at 16, 80, or 320 feet. Under Condition R, the comparison was placed at the remote distances and the standard nearby. As the distance of the standard was increased (Condition N) the size of the comparison had to be made progressively larger than the physical size of the standard in order to achieve apparent equality. As the distance of the comparison was increased (Condition R) it had to be made increasingly smaller in order to match the standard. At distances beyond 200 feet a comparison which was smaller than the physical size of the standard was required to produce apparent equality.

Increasing overestimation of size at distances of 20 feet and less has been demonstrated by Jenkin (1957, 1959). In his first experiment, Jenkin (1957) found that when the comparison was at 2 feet, it had to be made significantly larger than when it was at 10 feet in order to match the standard at 20 feet; i.e., apparent size increased significantly over the short distance interval from 2 to 10 feet. Since the average match at the near position exceeded the physical size of the standard, and at the far position was exceedingly close to the physical size of the standard, overestimation of size is indicated. This overestimation cannot be attributed to a space error because size judgments made with the variable at the same distance as the standard were not significantly different from the true size of the standard, while the difference between true and judged size was highly significant when the standard was at 20 feet and the variable at 2.

In order to study more fully the relationship between small increments of distance and estimates of objective size, Jenkin (1959) performed a second and a third experiment, in which he presented comparison stimuli at distances intermediate between those employed in his earlier study. In the second experiment, the comparison was located at a distance of 20, 40, or 160 inches. In the third experiment, a fourth distance (80 inches) was inserted between the 40- and 160-inch positions. For all distances, mean size matches exceeded the physical size of the standard stimulus and became significantly larger as the comparison object was placed closer to *O;* i.e., overestimation of size increased with distance. The use of a third and fourth comparison distance made it possible to plot the results graphically. When plotted against the logarithms of the distances, the mean size matches gave points fitted by a straight line. According to Jenkin (1959), this straight line relationship "suggests the existence of some hitherto undiscovered law relating apparent size and short increments in distance" (p. 348).

In his first experiment Jenkin used natural lighting. Coules (1955) has demonstrated that a brighter object farther away appears to be at the same distance as a nearer but dimmer object (see also Ittelson, 1952). If the more distant stimulus objects in Jenkin's experiments received relatively less illumination than the nearer objects, then progressive distance overestimation might have resulted. This in turn would account for the progressive overestimation of size which was obtained. In order to control for differences in illumination in his second experiment, Jenkin (1959) varied the illumination of the standard stimulus between 11 foot-candles and 26.5 foot-candles, while keeping the illumination of the comparison constant at 11 foot-candles. Differences in illumination of the standard had no significant effect either on amount of overestimation of size or on the rate at which it increased with distance.

In order to determine whether increasing overestimation of size would occur with a familiar stimulus object, Jenkin (1959) permitted *O* to examine the standard at a distance of 24 inches for 5 seconds before making size matches with the standard at its usual distance of 320 inches. Increased familiarity with the standard reduced the amount by which it was overestimated but did not affect the rate at which overestimation increased with distance.

In a further experiment Jenkin (1959) tested the possibility that decreasing size matches are related to decreasing ratios of distance between standard and comparison objects. This was accomplished by placing the standard 80 inch in front of *O* instead of 320 inches as formerly. If the distance ratio is crucial, then a steady decrement in the size match should be obtained from 20 to 80 inches, and an increment in the size match should be observed at 160 inches. The data of the third experiment did not confirm this expectation. The size matches decreased as the comparison receded from 80 to 160 inches.

From the experimental evidence which we have summarized, it appears that increasing overestimation of size is well-established. The invariance hypothesis demands that increasing overestimation of size be accompanied by a tendency for apparent distance to increase more rapidly than physical distance. At least one experiment indicates that apparent distance does increase in this way: Tada (1956) performed a bisection experiment in which secondary cues to distance were eliminated. Using binocular vision, *O* made bisections by stopping one of two light spots when it appeared to be halfway between *O* and the second spot, which was fixed at a point designating the total distance to be bisected. In a second experiment, *O*'s task was to bisect a 2- or 4-meter interval, presented at various distances from *O*, with each of its end points marked by a bright spot. In both experiments, Tada found that the phenomenal midpoint was farther than the objective one. In other words, the farther half of the distance was overestimated as compared with the nearer half.

Tada's findings are given some support by Purdy and Gibson (1955). They found that when *O* was permitted full primary and secondary cues to distance, errors in dividing distances (up to 300 yards) into halves and thirds tended most frequently to involve making the nearer segment too large in comparison with the farther. However, few errors were made; in general, perceived magnitudes of distance corresponded well with physical magnitudes of distance. Consistent findings of a large acceleration of apparent size with distance would seem to demand a reasonably large and consistent tendency to overestimate the farther distance as compared with the nearer.

The invariance hypothesis is further weakened by the fact that at least two experiments on distance estimation give results exactly opposite to those of Tada (1956). Gilinsky (1951) has presented evidence which indicates that perceived distance increases with true distance at a *diminishing* rate. The experimenter (*E*) moved a pointer at a slow and nearly constant rate along the ground away from *O*, who instructed *E* to mark off successive increments of equal perceived length. In the case of one *O*, every increment of apparent distance represented an attempt to match a memorized "subjective foot rule"; in the case of the other *O*, a memorized "subjective meter stick" was being matched. For both *O*s apparent distance increased more slowly than physical distance. This experiment is defective because error in bisection experiments is related to the direction of motion of the pointer; as the pointer withdraws, *O* tends to make the farther segment too large in comparison with the nearer (Purdy & Gibson, 1955). This defect was avoided in a second experiment by Gilinsky (1951). Across a large, flat lawn, a line was stretched perpendicular to the frontal, parallel plane of *O*. *O* was required to bisect each one of 14 distances, ranging from 8 to 200 feet, by stopping a pointer, which moved back and forth along the line, at a point which appeared to be halfway between the near end of the line and a marker indicating the total distance to be bisected. The results were the same as in the previous experiment.

Smith (1958) also found that far distances tend to be underestimated in comparison with near ones. As a standard stimulus he used a white sheet of oilcloth, which was spread out on the floor of a hall. The variable stimulus was a strip of the same oilcloth rolled from a small roller. To match the length of the standard, the variable was made 15.1% longer than the standard.

The invariance hypothesis must be

abandoned if we accept both the finding of apparent distance which increases less rapidly than physical distance and the finding of increasing overestimation of size. A way out of this difficulty is suggested by Carlson (1960b), who maintains that increasing overestimation of size is an artifact of "objective" instructions. When *O* is trying to judge actual, physical size, his size matches will be influenced by his beliefs about size-distance relationships. The major attitude by which *O* will be influenced is the concept of perspective—the notion that apparent size becomes smaller as distance increases. "From *O*'s point of view, a near object must 'look' larger than a far object for the two to be equal in physical size" (Carlson, 1960b, p. 200). Hence *O* will make size matches which appear to indicate an overestimation of the far object.

> Given several discriminably different distances in the same setting, amount of overestimation may be a fairly precise function of distance, but only because trials at different distances are not really independent, and *O* can judge the distances relative to each other (p. 201).

In support of his thesis, Carlson (1960b) pointed out that overestimation does not occur in experiments, such as those of Brunswik (1956, pp. 67–69) and Singer (1952), in which *O* is asked to base his size judgments upon a naive, natural impression of size ("look" instructions). Carlson (1960b) performed an experiment in which *O* was allowed, but not required to differentiate apparent visual size from objective size. Using free binocular regard, *O* adjusted a 10-foot distant variable triangle to match a 40-foot distant standard. *O* was also required to bisect the distance to the apparatus on which the standard triangle had been presented. Under apparent size instructions, size matches were accurate. Under objective size instructions, overestimation of size occurred. We are told that under both "look" and "objective" instructions, "the half-distance of the standard was . . . overestimated" (p. 206). Apparently this means that *O*, in bisecting the distance to the standard's apparatus, required the marker to be placed too close to himself. If so, the results indicate that apparent distance increases less rapidly than physical distance.

It is doubtful that Carlson has removed the difficulties facing the invariance hypothesis. Carlson (1960b) used only one pair of distances; if either the standard or the variable had been placed at more than one distance, he might have found that estimated size increases with distance under "look" instructions, even though overestimation does not occur. The published data of Brunswik (1956, pp. 67–69) and of Singer (1952) do not provide an answer to this question. Furthermore, Carlson's finding that size is accurately estimated does not match his finding that apparent distance increases less rapidly than physical distance.

Instead of linearly increasing overestimation of size, some investigators have reported a curvilinear relationship between physical distance and overestimation of size. Hastorf and Way (1952) found that when distance cues are available, overestimation of size increases from 10 to 20 feet and decreases from 20 to 30 feet. Chalmers (1952) found that overestimation increased from 10 to 20 feet and decreased from 20 to 50 feet when *O* viewed the 10-foot comparison binocularly.

It should be noted that even if the reported instances of progressive overestimation of size should be accounted for by progressive overestimation of distance this would still leave unexplained the curvilinear size distance relationship obtained by several investigators.

Nonmatching Judgments of Size and Distance

According to the invariance hypothesis, the perceived size of an object is proportional to its perceived distance, when its retinal image size is held constant. This requirement of proportionality is frequently not met when size and distance judgments are both made in the same experimental setting. For purposes of exposition, we may divide the experiments which produce nonproportional results into two classes: (*a*) In the first class are included those experiments which provide evidence for a "size-distance paradox"—a consistent tendency either to couple an underestimation of the relative size of an object with an overestimation of its relative distance or to couple an overestimation of the relative size of an object with an underestimation of its relative distance. (*b*) In the second class are those experiments which show that a variable having a consistent influence on size judgments has no consistent influence on distance judgments, and, correlatively, those experiments which show that a variable having a consistent influence on distance judgments has no consistent influence on size judgments.

Class 1: The Size-Distance Paradox. A striking example of the size-distance paradox is the moon illusion. As is well known, the moon appears larger on the horizon than at the zenith. According to the invariance hypothesis, it should also look farther away. Yet *O* usually reports that the moon looks closer when it is low in the sky. The most recent discussion of this time honored problem is by Kaufman and Rock (1960).

More detailed evidence for the size-distance paradox is provided in an experiment by Gruber (1954). The standard stimulus was a triangle which was alternately 10 and 15 centimeters in height. To the right of the standard was a variably sized triangle. This variable triangle was placed at six distances ranging from 200 to 450 centimeters. For each distance *O* made four kinds of judgments, all of them under "look" instructions:

1. *O* set the variable-size triangle so that it appeared equal in size to the standard (*a*) when the standard was half as far from *O* as the variable, and (*b*) when both stimulus objects were equidistant from *O*.

2. *O* adjusted the distance of the standard so that it appeared (*a*) half as distant as the variable, and (*b*) equidistant with the variable.

The results were all contradictory to the invariance hypothesis:

1. By setting the size of the variable significantly larger than the actual size of the standard in the size constancy matches (judgments of Type 1*a*), *O*s exhibited a mean *overestimation* of the relative size of the standard. However, a mean *underestimation* of the relative distance of the standard occurred; *O* set the standard sized triangle too far away in the half-distance judgments.

2. "Analysis of individual differences revealed no correlation between size and distance judgments." (Gruber. 1954. p. 426).

3. As the physical distance of the farther object increased, the mean constant error in size constancy matches rose from 4% to 23%, whereas the mean constant error in half-distance judgments did not vary progressively with absolute distance, fluctuating around 17%.

4. The mean errors in the control judgments (1*b* and 2*b*) were not large enough to account for the magnitude of the errors in the size

constancy and half-distance judgments.

By means of her size-distance equations, Gilinsky (1955b) attempted to show that Gruber's data are properly interpreted as supporting rather than rejecting the hypothesis that perceived size is proportional to perceived distance. However, Gruber (1956) has pointed out that Gilinsky's analysis does not apply to his most interesting result, the Finding 1 above that an object which is consistently overestimated in size is consistently underestimated in distance. Gilinsky's analysis deals only with Finding 3, and in order to do so, it must make use of a number of somewhat arbitrary assumptions.

Jenkin and Hyman (1959) report that when *O*s are given an "objective" set, Gruber's finding of a size-distance paradox is confirmed. Size judgments were obtained under four different distance conditions: (*a*) comparison 30 feet and standard 15 feet from *O*, (*b*) comparison 30 feet and standard 2 feet from *O*, (*c*) comparison 15 feet and standard 1 foot from *O*, and (*d*) comparison 15 feet and standard 15 feet from *O*. *O* made size judgments under two different instructions: to match for physical size, and to match for retinal image size. Following the size judgments, the black mounting-board upon which the variable had appeared was placed 30 feet from *O*, who was required to make estimates (in feet) of this distance. Under objective instructions, *O*s either judged the variable as relatively small and its mounting as relatively remote, or as relatively large and its mounting as relatively near.

> The relationship of analytic size-judgments to estimated distance was toward the distant object being regarded as relatively large and relatively remote, or relatively small and relatively near (p. 73).

Thus we have the paradoxical result that an *O* who is set to judge physical size responds as if he were ignoring the simple geometrical help which would come from taking distance into account, while a person who is deliberately trying to ignore distance in order to get retinal image matches responds as if he were taking distance into account. Assuming that the analytic judgments represented *O*'s best attempt to respond in terms of retinal image size, and assuming that objective size judgments represent perceived size, the invariance hypothesis demands, for any given distance, a positive correlation between analytic size judgments and objective size judgments. Such a correlation was not found.

Heinemann, Tulving, and Nachmias (1959) obtained nonmatching size and distance judgments in an experimental situation in which *O* was permitted only primary, monocular cues to distance. When distance judgments were being made, the comparison was held constant at 1° and *O* reported which of two successively presented disks, standard or comparison, was farther away. When judging the relative distance of a standard and a variable, most *O*s said that the objectively nearer disk was farther away. Since the far object looked nearer than the near object (which subtended the same retinal angle), it should have been judged as smaller than the near object, if the invariance hypothesis is true. Yet size matches were consistently "in the direction of size constancy"; the farther away an object was, the larger it was judged as being.

Kilpatrick and Ittelson (1953) have drawn attention to two phenomena of accommodation which involve a size-distance paradox. They cite Aubert's finding that partial paralysis of accommodation produces both a reduction in the apparent size of an object and an increase in its apparent distance. They also report von Kries' observation that an object appears to diminish in size and also to recede when *O* shifts fixation from that object to one closer by. However, both these findings are complicated by the fact that changes in accommodation involve changes in retinal image size (e.g., Pascal, 1952).

Nonproportional Results of Class 2. A number of studies indicate that when visual angle is constant, changes in apparent size are not consistently accompanied by changes in apparent distance, and changes in apparent distance are not consistently accompanied by changes in apparent size. Beginning with the classic experiments of Wheatstone and Judd, it has been frequently found that increases in convergence are regularly accompanied by decreases in apparent size. Insofar as the decrease in retinal image size accompanying convergent accommodation is not sufficient to account for the obtained decrease in apparent size, the invariance hypothesis requires that the decrease in apparent size be accompanied by a decrease in apparent distance. Yet the obtained changes in apparent distance are equivocal. This result was corroborated recently by Hermans (1954), who used a telestereoscope to produce six changes in convergence from 0 to 10°. As degree of convergence increased, the mean apparent size of the standard, as determined by *O*'s adjustment of a variable, decreased significantly. Verbal reports indicated that some *O*s perceived a decrease in distance with increasing convergence, while other *O*s perceived an increase in distance.

Kilpatrick and Ittelson (1953) found that an illusory movement in depth was not accompanied by the required change in apparent size. The trapezoidal window was suspended in *O*'s line of sight, so that its sides were vertical and the physically larger end of the trapezoid was farther from *O* than the smaller end. An ordinary playing card and a piece of cotton were successively moved through an opening in the window by means of a thread stretched at right angles to the line of sight. Objects carried through the trapezoid in a straight path by the moving thread appear to move through an S shaped path in the horizontal plane. In the majority of observation trials, *O*s perceived definite movement in depth of 2 feet. Yet for the largest number of trials on which movement in depth was perceived, no size changes were reported either for the playing card or for the cotton. On the remaining trials on which movement in depth was reported, size changes in a direction opposite to that required by the invariance hypothesis were reported about half as frequently as changes in the required direction. In a second experiment, an ordinary sized playing card was suspended from each of the two stationary wires by means of which the trapezoid was hung from the ceiling. On 19 trials *O*s perceived one card to be larger than the other. But on only 10 of these 19 trials did *O*s perceive the apparently larger card to be farther away, as required by the invariance hypothesis.

According to the invariance hypothesis, an improvement in *O*'s ability to estimate the distance of an unfamiliar object should result in an improvement in his ability to estimate its size. Using a series of photographs of the Gibson size-at-a-distance set-up (described above), Gibson and Smith (1952) found that training in estimation of the distances of the stakes in the photographs significantly improved *O*'s accuracy in estimating these distances. However, there was no significant improvement in *O*'s ability to estimate the sizes of the stakes.

Another finding contrary to the invariance hypothesis involves the visual tau effect (cf. Geldreich, 1934). Kilpatrick and Ittelson (1953) note that the difference in the perceived lateral separation of the points is not accompanied by any change in the apparent distance of the pairs of points from *O*.

Matching Judgments of Size and Distance. We have seen that most experiments which obtain size and distance judgments in the same setting provide evidence against the invariance hypothesis. However, in two experiments in which convergence provided the chief distance cue, matching size and distance judgments were obtained.

Bleything (1957) used a stereoprojector to cast two ring targets onto a screen. Observer and projector were equipped with polaroid filters making it possible for one ring to be seen with one eye only and the other ring to be seen with the other eye only. *O* saw a single fused ring which appeared to approach and recede in depth as *E* varied the distance between the center of the projected rings. As required by the invariance hypothesis, the apparent size of the fused ring increased with apparent distance, although the perceived size of the ring increased at a slightly greater rate than predicted by the formula, $s=(a)(d)$.

Roelofs and Zeeman (1957) report that when retinal image size is constant, a number of variables which affect apparent size also produce a complementary change in apparent distance. Two series of figures were presented. In the first series each card bore six figures: two pairs of equal sized circles which were fused binocularly (orthoptically) to give two perceived circles, and two circles which were presented either to the right or left eye alone. For the first series, Roelofs and Zeeman report the following findings:

1. Of the two circles seen binocularly, the one which required the greater convergence always appeared smaller. As required by the invariance hypothesis, it also always appeared closer to *O*.

2. The apparent size of the circles seen monocularly tended to be intermediate between the apparent sizes of the two circles seen binocularly. Matching this, the apparent distance of the monocularly seen circles tended to be intermediate between the apparent distances of the binocularly seen circles.

3. For a given card, the apparent size of a monocularly seen circle was closer to the apparent size of the binocularly seen circle from which it had the smallest physical separation on the card. As required, the apparent distance of the monocularly seen circle was also closer to the apparent distance of the nearby, binocularly seen circle.

4. The apparent size of the circles seen monocularly was just as strongly influenced by the circles seen binocularly with a weaker convergence as by the circles seen binocularly with a stronger convergence. However, the apparent distance was more strongly influenced by the circles seen with a stronger convergence. This is the only general finding of Roelofs and Zeeman which contradicts the invariance hypothesis.

5. Monocularly seen circles in the

lower half of the stimulus card tended to be perceived as smaller than and, matching this, as nearer than monocularly seen circles in the upper half of the card.

6. Monocularly seen circles in the nasal position tended to be seen as smaller than and as closer than monocularly seen circles in the temporal position.

7. Monocularly seen circles in the left half of the optic field tended to be seen as smaller than and as closer than monocularly seen circles in the right half of the field.

The second series of stimulus cards used by Roelofs and Zeeman had three equal sized circles: a single circle which was presented to one eye only and a pair of circles which were fused binocularly to give a single perceived circle. For the second series, the apparent size of the circles seen binocularly was greater than the apparent size of the circles seen monocularly. Matching this, the apparent distance of the binocular circles was greater than the apparent distance of the monocular circles. The findings obtained with the earlier stimulus series were corroborated with respect to the effects on apparent size and distance of nasal vs. temporal, right vs. left, and higher vs. lower stimulus positions.

Although the general findings of Roelofs and Zeeman are in accord with the invariance hypothesis, there were some individual exceptions to the required matching of size and distance judgments for all findings except the first.

In at least one respect, the experiments of Bleything (1957) and of Roelofs and Zeeman (1957) provide a fairer test of the invariance hypothesis than do the experiments which produce nonmatching judgments. Bleything, and Roelofs and Zeeman had *O* estimate the size and distance of the stimulus almost *simultaneously.* In the other experiments a relatively long temporal interval separated the estimations. It is possible that when *O* is asked to make adjustments of size (distance), his perception of size (distance) occupies the center of attention, and his perception of distance (size) is relegated to the background. The perception of both size and distance when they are merely registered as background may differ from their perception when they occupy the observer's close attention. Hence, when *O* is set to perceive size and distance at the same time, it is more likely that his judgments will match as required by the invariance hypothesis than when he is set to perceive only size or distance and not both.

Despite the methodological reservations mentioned immediately above there is sufficient cause for concluding that all is not well with the traditional formulation of the size-distance relationship. It remains to be seen whether the generally accepted invariance hypothesis can by any means be reconciled with the contradictory findings described in this section. In the eventuality that this reconciliation will prove impossible, then the way is open for a restatement of the size-distance relationship. It is also possible that in certain instances size and distance perception are unrelated. Despite their temporal co-occurrence these two experiences may be independent but simultaneous responses to separate aspects of the proximal stimulus situation. Some experimental evidence that this may indeed be the case has been presented by Gruber (1954) and Epstein (in press).

The Known Size-Apparent Distance Hypothesis

According to this hypothesis the known size of a stimulus object determines a unique relation of retinal image size to apparent distance. Two corollaries can be derived easily from this proposition:

Corollary 1. Discrete changes in the size of the retinal image of an object whose known size remains constant will be perceived as corresponding changes in the apparent distance of that object.

Every identified object may be said to possess an "assumed size." This term refers to "the entirely subjective sense of size which the observer might relate to a specifically characterized physiological stimulus-pattern" (Hastorf, 1950, p. 195). The second corollary deals with assumed size.

Corollary 2. Changes in the assumed size of an object whose retinal size remains constant will result in appropriate changes in the apparent distance of that object.

Corollary 1

Most of the investigations which have been reported are concerned with Corollary 1. An ingenious experimental test of this proposition which has been cited often was performed by Ittelson (1951). In one experiment three playing cards were presented singly to *O* under conditions of complete reduction. Each of the cards was placed at the same physical distance from *O*. The task for *O* was to adjust a comparison stimulus of familiar size, which was presented separately, until the comparison object and the standard playing card appeared to be at the same distance. The neat turn in this experiment concerns the sizes of the three cards: one was a normal sized card, all the dimensions of another one were doubled, and the dimensions of the third card were halved. Presumably, in this situation, the only cue available for the estimation of distance was retinal size which varies directly with changes in physical size when distance is constant. When known size is invariant, these changes in retinal size ought to be perceived as changes in distance and not as changes in size. The larger card should be localized at a point half way between *O* and the distance at which the normal card is perceived, and the smaller card should be localized at twice the distance of the normal card. The results for five *O*'s confirmed these expectations almost exactly (Ittelson, 1951, p. 64).

This experiment has been vigorously criticized by Hochberg and Hochberg (1952) who have argued that Ittelson and others have failed to distinguish between familiar size, on the one hand, and the *relative* size of the stimuli on the other (i.e., change or difference in size of objects of similar shapes). For this reason, Hochberg and Hochberg (1952) designed an experiment in which familiar size and relative size were separated. Two figures were drawn on a two-dimensional, reversible screen drawing. One panel contained a drawing of a man, and on the other panel a boy of the same size and approximate contour was represented. The question is whether the panel with the boy appears to be nearer more often than the panel containing the man. This is to be expected if familiar size is determining apparent localization. The results showed that familiar size was ineffective in this situation.

In a second experiment the effectiveness of *relative* size was tested. The same procedure was followed with one difference. Whereas the first experiment held relative size constant while familiar size was varied, the second experiment held familiar size constant while varying relative size. Both panels contained drawings of the same boy, but one was a reduced version of the other. Here, relative size would lead to localizing the panel containing the larger boy nearer than the other panel. The results were in agreement with this expectation. These findings led the authors to suggest that there may be a stimulus bound correlation between retinal size and perceived distance which would make the introduction of unconscious assumptions (about known size) unnecessary.

Further experimental evidence in support of this emphasis on relative size is presented by Hochberg and McAlister (1955). Four cards, each bearing one small figure and one large figure were presented singly. Card 1 bore a large circle and a small circle; Card 2, a large square and a small square; Card 3, a large circle and a small square; and Card 4, a large square and a small circle. In terms of relative size, it would be expected that Cards 1 and 2 should yield more three-dimensional responses than Cards 3 or 4. This was the case.

In a second experiment the authors inquired whether the direction of the three-dimensional responses is in accordance with what would be predicted in terms of relative size.

> In terms of the cue of relative size the larger figure should appear nearer than the small one in Cards 1 and 2. They did. If this were due to the operation of familiar size, we would expect similar results to hold with respect to Cards 3 and 4 (p. 296).

This did not happen.

Ittelson (1953) has replied to the above criticisms by citing several instances in which relative size is not involved. These are cases when only a *single* object is present in the field. Ittelson argues that if a single, familiar object viewed monocularly in a dark room is replaced by another of the same physical size, but of different assumed size, the apparent distance of the second will be different from the first. The clearest demonstrations of this effect have been Ames' "watch-card-magazine" experiment (1946–47) and Hastorf's similar investigations (1950). We will describe Hastorf's study later in this section when we consider Corollary 2.

In addition Ittelson (1953) maintains that if a single, familiar object is viewed monocularly in a dark room, it is perceived immediately and unequivocally at some definite distance which can be correctly predicted on the basis of the familiar size of the object. Finally, the claim is made that the size-distance perceptions related to a given stimulus can be changed by immediately prior experiences which change the size which is attributed to the stimulus. As an illustration Ittelson cites the experiments which demonstrate the influence of size assumptions on perceived radial motion (see Kilpatrick & Ittelson, 1951).

The latter two assertions are incompatible with an explanation based on the relative size cue. However, subsequent investigations have failed to confirm their validity, and have provided further support for the relative size thesis (also see Hochberg & Hochberg's—1953—rejoinder to Ittelson). The experiments reported by Gogel, Hartman, and Harker (1957) show that the retinal size of a familiar object is totally inadequate as a cue for the *absolute* apparent distance of that object. The investigations reported by Epstein (in press) confirm the findings of Gogel et al. and also demonstrate that experiences which modify *O*s assumptions concerning object size do not modify his perceptual experience. The problem for Gogel et al. (1957) was to "investigate whether the retinal subtense of a familiar object can act as a determiner of the apparent *absolute* distance of that object from the observer" (p. 1). This study employed a nonvisual method of measuring perceived distance of the object. *O* was asked to throw a dart to the perceived distance without seeing the results of the throw. Since successive throws might involve relative distance judgment, only the response to the object which was *first* perceived was considered in measuring the perceived absolute distance of that object. The stimulus object was a normal or double sized playing card, located at a distance of 10 or 20 feet in a reduced cue situation.

The distance responses for the stimuli initially presented did not confirm the expectations which follow from the Known Size-Apparent Distance Hypothesis. Not only did the results fail to agree with any precise predictions of apparent localization, e.g., the double sized card at a physical distance of 20 feet should be localized at 10 feet, but the less stringent prediction, e.g., the double sized card should appear to be nearer than the normal card, was also not confirmed. Under these conditions perceived distance was totally unrelated to retinal size.

When a similar analysis was performed for all of the four reduced cue situations collectively (i.e., the *same* *O*s in all four situations), partial support was obtained for the Known Size-Apparent Distance Hypothesis in its less precise formulation. The implication of this finding is clear. The secondary analysis shows only that *relative* distance perception, as some function of *relative* retinal subtense, can occur for successively presented stimuli.

The first of three experiments reported by Epstein (in press) was essentially a replication of Ittelson's (1951) experiment with two major modifications: (*a*) prior to the judgmental task *O*s in the Experimental Group participated in a card game which was designed to modify their assumption concerning the normal size of cards, and the constancy of the physical size of cards, (*b*) at the conclusion of the distance settings all *O*s were required to judge the apparent size of the stimuli.

The results of this experiment did not support the known size hypothesis. Despite the modifying treatment experienced by the Experimental Group there was no difference between the distance judgments of the Experimental Group and a Control Group which did not have prior training. In addition, none of the distance judgments met the precise quantitative requirements of the known size thesis, e.g., while the quarter sized card appeared to be more distantly located than the normal card, it was *not* set at four times the distance of the normal card. Finally, the stimuli of different physical size were also judged to be of different size.

In Experiment II it was demonstrated that similar apparent distance effects would obtain when only relative retinal size is operative (known size and assumed constancy of physical size absent). Finally, in Experiment III it was shown that in the absence of the relative size cue no systematic size-distance effects are obtained. The results of Experiments II and III bolster the position adopted by Hochberg and Gogel.

In this connection the results reported by Gogel and Harker (1955) may also be cited. Gogel and Harker obtained judgments of apparent distance for two playing cards of different sizes under reduced cue and near complete cue conditions. They found that the relative apparent depth of the two cards was a function of the lateral separation between the two cards. They concluded that "the effectiveness of size cues to relative depth increased as the lateral separation of the differently sized cards was increased" (p. 315). There is no reason to expect such results if the original depth effects were based on the operation of an assumed size factor.

This review leads to the conclusion that despite its reasonableness Corollary 1 of the Known Size-Apparent Distance Hypothesis is unnecessary. Many of the experimental effects which are most frequently cited as evidence for its validity are more simply attributed to other factors, e.g., relative size. In those cases in which these factors are eliminated the "Known Size Effect" is also eliminated. The question remains whether all reported effects of known size on apparent localization can be explained in this way. This brings us to Corollary 2 of the Known Size-Apparent Distance Hypothesis.

Corollary 2

The second corollary requires that a *single* object whose physical size remains unaltered will undergo changes in apparent spatial localization with changes in the physical size which *O* attributes to the object. Thus, if the same object is assumed by *O* to have a small size at one time, and a large assumed size at a later time, it will be perceived to be more distant at this later time although the physical distance of the object is the same at both times. It is obvious that effects of this nature cannot be accounted for by processes which depend on the opportunity for comparisons of successively presented stimuli which differ along a physical dimension.

There are very few experimental studies which demonstrate that such an effect does indeed obtain. In Hastorf's (1950) investigations a rectangular or circular area of light was given a "large assumed size meaning" or a "small assumed size meaning." That is, the rectangle was called either an envelope or a calling card, and the circle was called either a billiard ball or a ping-pong ball. The size at which the stimulus was set, in order to appear at a specific distance, varied when the assumed size attributed to the stimulus was varied by the size suggestion, i.e., by naming the stimulus.

In a study of the effects of past experience on apparent size, Smith (1952) reported findings which may be interpreted in the same way. In the first stage of the experiment *O* judged the apparent distance of several simple geometrical forms, e.g., circles and squares. Then, over a period of 2.5 weeks *O*s participated in a series of tasks requiring the manipulation and discrimination of geometrical forms of the same shape but larger in size. In this way *E* hoped to alter the attributed size of the original forms. Then the *O*s were retested, i.e., *O*s repeated the judgments which were made prior to training. The distance judgments were observed to change in the direction demanded by the modification in attributed size.

Finally some incidental findings of Ittelson (1951) may be mentioned. In one variation of the experiment described earlier *O* judged the apparent distance of a half sized playing card and a matchbox of identical size when both were located at the same objective distance of 7.5 feet. The playing card was localized at a distance of 14.99 feet while the matchbox was judged to be at a distance of 8.96 feet. Apparent distance was influenced by *O*s assumptions concerning the physical size of the stimulus objects.

Corollary 2 has received support from the investigations described above. Still, there is clearly a need for further experimentation. In particular it would be useful to have the results of experiments which meet the following three requirements:

1. A measure of *O*'s *immediate* perceptual impression should be obtained. In most cases *O* has been allowed an extended period of time in which to make an adjustment which he is "satisfied with." Under such conditions many judgmental and attitudinal factors may enter into the adjustment process, and contaminate or at least alter the identity of the effect.

2. Different *O*s should be used for the various attributed size conditions. It is possible that the same *O* performing under the various conditions may be making memorial comparisons between the first attributed size-apparent distance judgment and the requirements of the current situation. This possibility is minimized if an extended temporal interval intervenes between the required judgments. Nonetheless, even though 6 days intervened between successive critical judgments in Hastorf's experiments, Hastorf (1950) reports that "some subjects did appreciate the fact that it was the same stimulus objects being given two different names" (p. 208).

3. In addition to these two requirements it might be helpful to obtain a measure of apparent size independently of *O*'s distance judgments. The results of earlier experiments suggest that such information may be instructive.

The Relationship between the Size of the Afterimage and Distance

A special case of invariance is Emmert's Law. The law states that the size of a projected afterimage (AI) is directly proportional to the distance from the eye to the projection surface. This statement follows from simple geometric considerations if we keep in mind that for the case of AIs the subtended visual angle remains constant regardless of variations in projection distance. The apparent simplicity disappeared following Boring's (1940) well-known attempt to demonstrate that Emmert's Law implies its converse, size constancy. Boring's thesis has been expressed succintly by Edwards (1950):

> What Boring was saying was that apparent size must increase with constant retinal size and increasing distance, if it is also true that apparent size remains constant with shrinking retinal size and increasing distance (p. 611).

We will not review the logic of Boring's formulations. It will suffice to point out that these formulations hinge on Boring's substitution of apparent size for physical size in the optical geometry of Emmert's Law. This substitution has been strongly criticized by Young (1950). Nevertheless, Boring's thesis has stimulated the major portion of writings concerned with Emmert's Law in the last 10 years. This work has followed two main themes.

The Historical Issue

Young (1950, 1951) has contended that Emmert intended to deal only with nonpsychological, Euclidian optical relationships. The contention is that Emmert's original reference (1881) was to the physical size of the

AI as determined by direct physical measurement of the occluded area on the projection surface. Young also maintains that a fundamental difference exists between real objects and AIs, and that it is inappropriate to speak of the apparent size of the latter.

The opposing view holds that Emmert was either concerned directly with apparent size and had, himself, implicitly made the substitution of *s* for *p* for the special case of AIs (Edwards, 1950), or that he did not distinguish the two different meanings of size perception (Boring & Edwards, 1951). The determination of apparent size requires a comparative technique. This method usually takes the form of judging the size of the critical object on the basis of an adjustable comparison stimulus or a series of different sized.stimuli. These, generally, are separated both in the lateral and frontal plane from the critical object. This method has found wide application in research on size constancy where apparent size is the crucial dimension.

Despite a careful reading of Emmert's original article (1881) there is little that we can contribute toward a resolution of this historical issue.[4] The one experiment which Emmert described in detail did utilize comparative stimuli, but both were attached directly to the projection surface. We are inclined to agree with Boring and Edwards (1951) that Emmert, in his own research, was not making a clear distinction between physical and apparent size.

The Theoretical Issue

The second aspect of the controversy is of greater significance. If Emmert's Law and size constancy are derivable from the same processes, then those conditions which determine the perceived size of real objects should affect the size of the AI also. If communality of process is not the case then the size of the AI should be unaffected by the same variables which affect the perceived size of real objects (or at least the effects should not be identical).

Edwards (1950) suggested that an experimental decision on this matter depends in part on the selection of an appropriate method of measurement. *E* can adopt either of two methods: (*a*) indirect measurement by employing a comparison stimulus or (*b*) direct measurement on the plane of projection. Edwards predicted that under reduced cue conditions Emmert's Law would fail when measured by Method *a* (i.e., the size of the AI would remain constant with increasing distance) but would hold when measured by Method *b*. Much of Edwards' position had been stated earlier by Helson (1936). In this paper Helson interpreted his results as showing that:

> when cues to distance and surroundings are eliminated the apparent size remains practically constant while the measured size of the projected image tends to obey Emmert's Law (p. 638).

Edwards (1953) tested one aspect of this prediction, viz., that the apparent size of the AI when measured by the comparison method would not conform to Emmert's Law under reduced cue conditions. *O*s projected AIs monocularly on to a dimly illuminated screen while looking through a reduction tube. The distance of the projection screen varied in five steps from 42 to 90 inches. A 2-inch luminous square in the same reduced field was adjusted until it appeared equal in size to the AI. No significant differences between the various dis-

[4] We are indebted to Martin Scheerer of the Department of Psychology of the University of Kansas for his expert translation of Emmert's article.

tances were obtained. Edwards concluded that Emmert's Law (i.e., "Emmert's Law of Apparent Size") had failed under reduction conditions. However, as Edwards himself admits, it is somewhat tenuous to uphold a prediction on the basis of confirmation of the null hypothesis.

Hastorf and Kennedy (1957) also contend that the controversy concerning the relationship of Emmert's Law to size constancy is primarily a matter of the type of measurement used. Under reduction and nonreduction conditions *O*s judged the size the real objects and AIs at various distances by the comparison method and the direct method (bracketing spotlights). The results for the comparison method confirmed Edwards' position, i.e., in the reduced cue situation, size constancy was greatly decreased and Emmert's Law did not obtain. With direct measurement there was no significant difference in the size of the AI between the reduced and full cue situations. This outcome supports Young's position. Thus, both sides of the controversy received support as did the authors' contention that the controversy hinges on different measurment techniques. However, Hastorf and Kennedy also reported that the use of bracketing spotlights in a dark room might provide a distance cue. If this is true, then it must be concluded that the direct measurement of the physical size of the AI under authentic reduction conditions remains to be accomplished.

Crookes (1959) takes a somewhat different approach to the problem of measurement. Crookes agrees with Young (1950, 1951) that Emmert's Law concerns "real," not apparent size. Further, he proposes that if Boring (1940) is right, Emmert's Law and size constancy should hold equally well when apparent size matches are obtained under the same conditions. Using the comparison method under "analytical" instructions, i.e., stressing retinal size, *O*s matched AIs and real objects. Crookes found that *O* made significantly better matches (i.e., showed significantly more constancy) in the case of the real objects. Crookes concludes that the subsumption of Emmert's Law and size constancy under a common heading is not justified. However, the objection could be raised that the analytical attitude induced by the instructions does not suit the purposes of research on constancy phenomena. Also, there is some question whether the greater constancy in the case of the real objects might not be due to the relatively greater ease of viewing real objects.

These studies concerned with the relationship between Emmert's Law and size constancy are not unanimous in their conclusions. Nevertheless, it is generally conceded that the method of measuring the AI may be critical. Thus, we might expect two or more forms of Emmert's Law to emerge, each embodying its own mode of measurement and each bearing a different relationship to other size-distance phenomena.

New approaches to measurement should be tried in this context, especially those promising some increase in precision. For example, Onizawa (1954) has developed a method whereby a screen bearing a comparison stimulus moves away from *O*, while a projection screen bearing an AI moves toward *O*. When *O* perceives equality between the AI and the comparison stimulus, he stops this movement. Ratios based on the respective distances of the two screens from *O* are compared with like ratios predicted from Emmert's Law. Onizawa presents data which indicate that his technique incurs less variability than the method of directly measuring the AI on the projection surface.

However, before the role of different measures can be clearly evaluated it will be necessary to test them together under identical conditions (e.g., reduction conditions). This requires that a given measure must not, itself, disqualify such conditions. Hastorf and Kennedy's (1957) observation (i.e., spotlights provide distance cues under reduced conditions) illustrates this problem.

Another matter deserving comment is related to the hybrid nature of Boring's formulation. While Boring has substituted apparent size for real size he has not seen fit to substitute apparent distance for physical distance. A careful reading of Boring's discussion (1942, p. 292) reveals a confusion of physical distance with apparent distance. The two terms are used interchangeably, seemingly without regard for any differences which may exist. It would be interesting to obtain pairings of the apparent size of the AI with the apparent distance of the projection surface. Such relationships if found to conform to Emmert's Law could hardly be explained in terms of the requirements of Euclidian geometry which applies only to physical distances and extents.

In this regard an additional complicating factor has been described by Ohwaki (1955). While expected values of Emmert's Law have been based on retinal size arising from the physical size and distance of the fixation object, Ohwaki (1955) found that perceived, not physical, distance was crucial in determining retinal size. Perceived distance was effective with either ordinary distance cues or past experience available. The interpretation was offered that it is perceived distance which underlies accommodation. Accommodation in turn regulates the size of the retinal image.

Finally, the problem of the physical as opposed to apparent size of the fixation object should be mentioned. Although this problem has received recent treatment in studies of figural aftereffects, its relevance with respect to Emmert's Law has not been explored.

It seems obvious that a refined statement of Emmert's Law must await intensive treatment of the variables discussed above (i.e., apparent distance of the projection surface, apparent size and distance of the fixation figure).

Other Determinants of the Size of the Afterimage

In a series of experiments, Young investigated the effect of a number of additional variables on the size of projected AIs using spotlights to outline the AI on the projection plane. In one study Young (1952a) varied the exposure time of the stimulus object in seven steps ranging from 0.01 to 40.0 seconds. No significant variations in the size of the AI were found with variations in stimulation time. Young (1952b) also investigated several features of the projection ground. In one experiment the illumination on the projection ground was varied through five log steps. No variation in the size of the AI was found. Another experiment (1952b) utilized pictures containing strong linear perspective. AIs were projected to specified points on these pictures and compared with AIs projected to similar points on a blank screen. The surfaces with linear perspective were found to influence AI size. It is tempting to account for these results by referring to presumed changes in apparent distance resulting from the differences in geometric perspective. Unfortunately this interpretation is complicated by the finding that there was little agreement between the *O*s in degree or direction of the size effect. However, an earlier study by Frank reported by Koffka (1935, p. 212) lends credence to an apparent distance interpretation. Frank used a perspective drawing of a deep tunnel. AIs projected to a phenomenally remote part of this tunnel were considerably larger than those projected to a near part. A similar effect is observed in the "Afterimage Demonstration" (Ittelson, 1952, pp. 32–33). Appropriate adjustments of the interposition indications using the overlay demonstration apparatus (Ittelson, 1952, p. 13) produce changes in the apparent distance of the projection surface and proportional changes in the apparent size of the AI.

The final study in this series (Young, 1952c) concerned the effect of large distances. In daylight AIs were projected in an open field to distances ranging from 25 to 1,250 meters. In each case obtained values were less than those expected on the basis of Emmert's Law. The hypothesis was advanced that with a brighter fixation stimulus (a square with a luminance of approximately 1,700 millilamberts), the retinal image is smaller, and consequently, the AI is smaller.

An interesting sidelight to the type of research on Emmert's Law considered so far is Oswald's (1957) study of the peripheral and central origins of AIs. Oswald uses these terms to contrast AIs in which the stimulation is confined to the retina with those involving the higher "representative" or brain centers. He cites a number of investigations, including his own in which AIs were obtained peripherally by presenting a light to an eye temporarily blinded by local pressure to the eyeball. Oswald also reviews a number of positive and negative reports of "central" AIs following imagined (visualized) objects or objects experienced in dreams. In his main experiment *O*s "imagined" crosses or squares and then projected AIs to a screen at various distances. Most *O*s were able to achieve AIs to imagined stimuli. However, very few AIs conformed to Emmert's Law. In this regard Oswald cites several earlier reports that eidetic *O*s deviate markedly from Emmert's Law when real stimuli are employed.

With further reference to individual differences, Brengelman (1956) found deviations from Emmert's Law to be larger in his neurotic group than with normals and psychotics.

Both large individual differences, such as those reported by Oswald, as well as smaller but consistent ones are inexplicable from the standpoint of a purely physical law. As an example of the latter kind, Young (1948) reported that all of his *O*s ($N=5$) yielded values falling consistently short of Emmert's Law values by a small margin. One would expect that variations due to inaccuracies of measurement alone would be randomly distributed.

Concluding Discussion

It seems to us that at least one compelling conclusion emerges from the survey we have just completed: the size-distance relationship expressed in the several formulations of the invariance hypothesis should not be assigned a unique or primary status in explanations of space perception. We have seen that this is only one of the several possible and actual relationships which are obtained. This need not cause any great consternation to those who recall the origin of the hypothesis in Euclidean geometrical principles. Although the distinction is sometimes overlooked it should be clear that the invariance hypothesis is a psychological proposition, and not a geometrical proposition. By no stretch of the imagination can Euclid's principles be applied *directly* to space perception. Of course, the analogy is plain and very tempting, and a successful translation would have been a happy logical circumstance. Nonetheless, failure to accomplish this translation should not cause surprise.

This brings us to a second remark. A great deal of logical and experimental analysis has been aimed at clarifying the term "size." We now distinguish not only real physical size, apparent size, and retinal size, but also assumed size, apparent angular size, etc. Usually the investigator makes explicit which aspect of size perception he is dealing with. However, with regard to distance, there is often a confusion of physical distance and apparent distance. We have seen that there is no unequivocal 1:1 relationship between physical distance and apparent distance. Therefore, it is not clear how experimental investigations of the size-distance relationship are to be interpreted when apparent distance judgments are not obtained. It seems to us that all studies of size and distance should obtain paired size-distance judgments.

This brings us to the methodological point which we mentioned earlier. Almost all of the experiments which have obtained paired size-distance judgments (including Epstein, in press) have done so in a successive judgment situation. We have already indicated the reasons for our dissatisfaction with this procedure. Here we wish only to reiterate the desirability for future investigation which employs a simultaneous judgment technique.

Finally, we wish to endorse a comment made earlier by Kilpatrick and Ittelson (1953) concerning individual differences. In order to assess the generality of the various size-distance hypotheses we need to look more carefully at the results of the performances of individual *O*s. In repeating some of the published research the first author has often been struck by the degree of interobserver and intraobserver variability. Results confirming various aspects of the invariance hypothesis do not allow *E* to say much about the individual *O*. In view of the "lawfulness" which is usually ascribed to the invariance hypotheses this extreme variability cannot be overlooked.

REFERENCES

Adler, F. H. *Gifford's textbook of opthalmology.* (7th ed.) Philadelphia: Saunders, 1959.

Allen, M. J. An investigation of the time characteristics of accommodation and convergence of the eyes. *Amer. J. Optom.*, 1953, **30**, 393–402.

Ames, A., Jr. *Nature and origin of perception: Preliminary laboratory manual for use with demonstrations disclosing phenomena which increase our understanding of the nature of perception.* Hanover: Inst. Ass. Res., 1946–47.

Bartley, S. H. *Principles of perception.* New York: Harper, 1958.

Bedrossian, E. H. *The eye: A clinical and basic science book.* Springfield: Charles C Thomas, 1958.

Bleything, W. B. Factors influencing stereoscopic localization. *Amer. J. Optom.*, 1957, **34**, 416–429.

Boring, E. G. Size-constancy and Emmert's law. *Amer. J. Psychol.*, 1940, **53**, 293–295.

Boring, E. G. *Sensation and perception in the history of experimental psychology.* New York: Appleton-Century, 1942.

Boring, E. G., & Edwards, W. What is Emmert's law. *Amer. J. Psychol.*, 1951, **64**, 416.

Brengelmann, J. C. Duration, periodicity, distortion and size of the negative visual afterimage in neurosis and psychosis. *Psychol. Beit.*, 1956, **2**, 569–585.

Brunswik, E. *Perception and representative* *design of psychological experiments.* Berkeley: Univer. California Press, 1956.

CARLSON, V. R. Apparent size, apparent distance, and real confusion. Paper read at American Psychological Association, Chicago, September 1960. (Complete text obtainable from Convention Reports Duplicating Service, Cleveland.) (a)

CARLSON, V. R. Overestimation in size-constancy judgments. *Amer. J. Psychol.*, 1960, **73**, 199–213. (b)

CHALMERS, E. L. Monocular and binocular cues in the perception of size and distance. *Amer. J. Psychol.*, 1952, **65**, 415–423.

CHALMERS, E. L. The role of brightness in primary size-distance perception. *Amer. J. Psychol.*, 1953, **66**, 584–592.

COMALLI, P. E., JR. The effect of time on distance-perception. *Clark U. Bull.*, 1951, **23**(203), 139–140. (Abstract)

COULES, J. Effect of photometric brightness on judgments of distance. *J. exp. Psychol.*, 1955, **50**, 19–25.

CROOKES, T. G. The apparent size of after-images. *Amer. J. Psychol.*, 1959, **72**, 547–553.

DEMBER, W. N. *Psychology of perception.* New York: Holt, 1960.

EDWARDS, W. Emmert's law and Euclid's optics. *Amer. J. Psychol.*, 1950, **63**, 607–612.

EDWARDS, W. Apparent size of after-images under conditions of reduction. *Amer. J. Psychol.*, 1953, **66**, 449–455.

EMMERT, E. Grössenverhältnisse der Nachbilder. *Klin. Mbl. Augenheilk.*, 1881, **19**, 443–450.

EPSTEIN, W. The known size-apparent distance hypothesis. *Amer. J. Psychol.*, in press.

GELDREICH, E. W. A lecture-room demonstration of the visual tau effect. *Amer. J. Psychol.*, 1934, **46**, 483–485.

GIBSON, E. J., & SMITH, J. The effect of training in distance estimation on the judgment of size-at-a-distance. *USAF Hum. Resour. Res. Cent. res. Bull.*, 1952, No. 52–39.

GIBSON, J. J. (Ed.) Motion picture testing and research. *AAF Aviat. Psychol. res. Rep.*, 1947, No. 7.

GIBSON, J. J. *The perception of the visual world.* Boston: Houghton Mifflin, 1950.

GILINSKY, A. S. Perceived size and distance in visual space. *Psychol. Rev.*, 1951, **58**, 460–482.

GILINSKY, A. S. The effect of attitude upon the perception of size. *Amer. J. Psychol.*, 1955, **68**, 173–192. (a)

GILINSKY, A. S. The relation of perceived size to perceived distance: An analysis of Gruber's data. *Amer. J. Psychol.*, 1955, **68**, 476–480. (b)

GOGEL, W. C. Relative visual directions as a factor in depth perceptions in complex situations. *USA Med. Res. Lab. Rep.*, 1954, No. 148.

GOGEL, W. C. Relative visual direction as a factor in relative distance perceptions. *Psychol. Monogr.*, 1956, **70**(11, Whole No. 418). (a)

GOGEL, W. C. The tendency to see objects as equidistant and its inverse relation to lateral separation. *Psychol. Monogr.*, 1956, **70**(4, Whole No. 411). (b)

GOGEL, W. C., & HARKER, G. S. The effectiveness of size cues to relative distance as a function of lateral visual separation. *J. exp. Psychol.*, 1955, **50**, 309–315.

GOGEL, W. C., Hartman, B. O., & HARKER, G. S. The retinal size of a familiar object as a determiner of apparent distance. *Psychol. Monogr.*, 1957, **71**(13, Whole No. 442).

GRUBER, H. E. The relation of perceived size to perceived distance. *Amer. J. Psychol.*, 1954, **67**, 411–426.

GRUBER, H. E. The size-distance paradox: A reply to Gilinsky. *Amer. J. Psychol.*, 1956, **69**, 469–476.

GULICK, W. L., & STAKE, R. E. The effect of time on size-constancy. *Amer. J. Psychol.*, 1957, **70**, 267–279.

HASTORF, A. H. The influence of suggestion on the relation between stimulus size and perceived distance. *J. Psychol.*, 1950, **29**, 195–217.

HASTORF, A. H., & KENNEDY, J. L. Emmert's law and size-constancy. *Amer. J. Psychol.*, 1957, **70**, 114–116.

HASTORF, A. H., & WAY, K. S. Apparent size with and without distance cues. *J. gen. Psychol.*, 1952, **47**, 181–188.

HEINEMANN, E. G., TULVING, E., & NACHMIAS, J. The effect of oculomotor adjustments on apparent size. *Amer. J. Psychol.*, 1959, **72**, 32–45.

HELSON, H. Size constancy of projected after images. *Amer. J. Psychol.*, 1936, **48**, 638–42.

HERMANS, T. G. The relationship of convergence and elevation changes to judgments of size. *J. exp. Psychol.*, 1954, **48**, 204–208.

HOCHBERG, C. B., & HOCHBERG, J. E. Familiar size and the perception of depth. *J. Psychol.*, 1952, **34**, 107–114.

HOCHBERG, C. B., & HOCHBERG, J. E. Familiar size and subception in perceived depth. *J. Psychol.*, 1953, **36**, 341–345.

HOCHBERG, J. E., & MCALISTER, E. Relative size vs. familiar size in the perception of represented depth. *Amer. J. Psychol.*, 1955, **68**, 294–296.

Holway, A. H., & Boring, E. G. Determinants of apparent visual size with distance variant. *Amer. J. Psychol.*, 1941, **54**, 21–37.

Howarth, E. The role of depth of focus in depth perception. *Brit. J. Psychol.*, 1951, **42**, 11–20.

Ittelson, W. H. Size as a cue to distance. *Amer. J. Psychol.*, 1951, **64**, 188–202.

Ittelson, W. H. *The Ames demonstrations in perception.* Princeton: Princeton Univer. Press, 1952.

Ittelson, W. H. Familiar size and the perception of depth. *J. Psychol.*, 1953, **35**, 235–240.

Jenkin, N. Effects of varied distance on short-range size judgments. *J. exp. Psychol.*, 1957, **54**, 327–331.

Jenkin, N. A relationship between increments of distance and estimates of objective size. *Amer. J. Psychol.*, 1959, **72**, 345–364.

Jenkin, N., & Hyman, R. Attitude and distance-estimation as variables in size-matching. *Amer. J. Psychol.*, 1959, **72**, 68–77.

Judd, C. M. An optical illusion. *Psychol. Rev.*, 1898, **5**, 286–294.

Kaufman, L., & Rock, I. The moon illusion: The problem reopened. Paper delivered at Eastern Psychological Association, Philadelphia, April 1960.

Kilpatrick, F. P., & Ittelson, W. H. Three demonstrations involving the visual perception of movement. *J. exp. Psychol.*, 1951, **42**, 394–402.

Kilpatrick, F. P., & Ittelson, W. H. The size-distance invariance hypothesis. *Psychol. Rev.*, 1953, **60**, 223–231.

Koffka, K. *Principles of gestalt psychology.* New York: Harcourt, Brace, 1935.

Leibowitz, H., Chinetti, P., & Sidowski, J. Exposure duration as a variable in perceptual constancy. *Science*, 1956, **123**, 668–669.

Lichten, W., & Lurie, S. A new technique for the study of perceived size. *Amer. J. Psychol.*, 1950, **63**, 281–282.

Ohwaki, S. On the factors determining accommodation: Research on size constancy phenomenon. *Tohuku psychol. Folia*, 1955, **14**, 147–158.

Onizawa, T. Research on the size of the projected after-image: I. On the method of measurement. *Tohuku psychol. Folia*, 1954, **14**, 75–78.

Oswald, I. After images from retina and brain. *Quart. J. exp. Psychol.*, 1957, **9**, 88–100.

Pascal, J. I. Effect of accommodation on the retinal image. *Brit. J. Ophthalmol.*, 1952, **36**, 676–678.

Purdy, J., & Gibson, E. J. Distance judgment by the method of fractionation. *J. exp. Psychol.*, 1955, **50**, 374–380.

Renshaw, S. Object perceived-size as a function of distance. *Optom. Wkly.*, 1953, **44**, 2037–2040.

Roelofs, C. O., & Zeeman, W. P. C. Apparent size and apparent distance in binocular and monocular distance. *Ophthalmologica*, 1957, **133**, 188–204.

Singer, J. L. Personal and environmental determinants of perception in a size constancy experiment. *J. exp. Psychol.*, 1952, **43**, 420–427.

Smith, O. W. Distance constancy. *J. exp. Psychol.*, 1958, **55**, 388–389.

Smith, W. M. Past experience and the perception of visual size. *Amer. J. Psychol.*, 1952, **65**, 389–403.

Smith, W. M. A methodological study of size-distance perception. *J. Psychol.*, 1953, **35**, 143–153.

Tada, H. Overestimation of farther distance in depth perception. *Jap. J. Psychol.*, 1956, **27**, 204–208.

Vernon, M. D. *A further study of visual perception.* Cambridge: Cambridge Univer. Press, 1954.

Wallach, H., & McKenna, V. V. On size-perception in the absence of cues for distance. *Amer. J. Psychol.*, 1960, **73**, 458–460.

Wohlwill, J. F. Developmental studies of perception. *Psychol. Bull.*, 1960, **57**, 249–288.

Woodworth, R. S., & Schlosberg, H. *Experimental psychology.* (Rev. ed.) New York: Holt, 1954.

Young, F. A. The projection of after images and Emmert's law. *J. gen. Psychol.*, 1948, **39**, 161–166.

Young, F. A. Boring's interpretation of Emmert's law. *Amer. J. Psychol.*, 1950, **63**, 277–280.

Young, F. A. Concerning Emmert's law. *Amer. J. Psychol.*, 1951, **64**, 124–128.

Young, F. A. Studies of the projected after images: I. Methodology and the influence of varying stimulation times. *J. gen. Psychol.*, 1952, **46**, 73–86. (a)

Young, F. A. Studies of the projected after images: II. The projection ground and the projected image. *J. gen. Psychol.*, 1952, **47**, 195–205. (b)

Young, F. A. Studies of the projected after images: III. The projection over large distances. *J. gen. Psychol.*, 1952, **47**, 207–212. (c)

Zeigler, P., & Leibowitz, H. Apparent visual size as a function of distance for children and adults. *Amer. J. Psychol.*, 1957, **70**, 106–109.

(Received September 21, 1960)